CREDIT EDA CASE STUDY

Prepared By

Purvi Gupta

Introduction & Objectives

- The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.
- This case study aims to give an idea of applying EDA in a real business scenario and will also cover risk analytics in banking and financial services and showcase how data is used to minimize the risk of losing money while lending to customers.

Problem Statement

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company. Identify patterns which indicate if applicants are capable of repaying the loan and their application should not be rejected.
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company. Identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

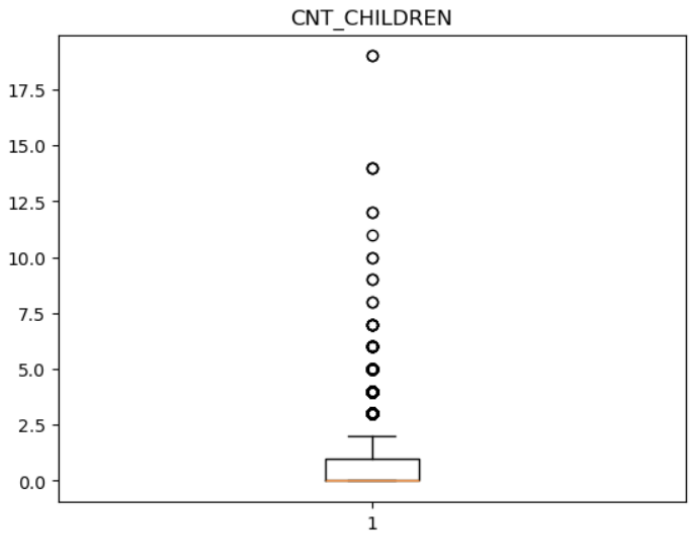
Identification of such applicants to eliminate the above risks using EDA is the aim of this case study.

Analysis and steps taken :

1. Data sourcing and loading datasets
2. Data understanding
3. Data cleaning –
    - fixing row and columns,
    - Imputing and removing missing columns,
4. Check for data imbalance
5. Identify outliers
6. Perform univariate analysis and bivariate analysis
7. Merge current and previous application data set
8. Analysis of merged dataset.

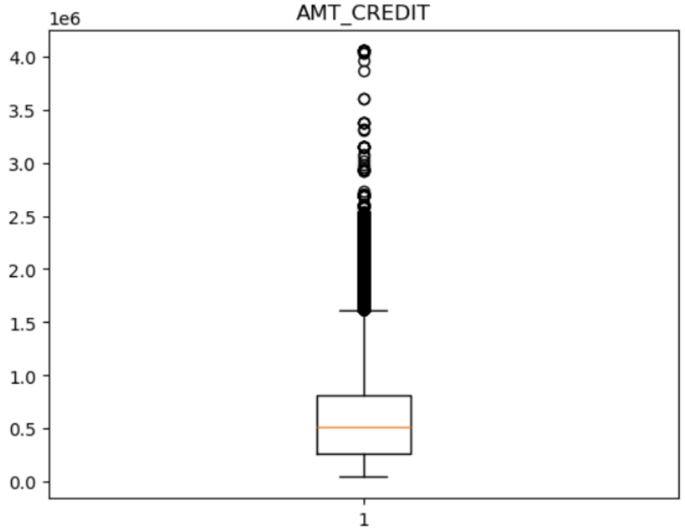
Outliers

Outliers are data points that deviate significantly from the majority of the dataset, often indicating anomalies, variability, or errors.



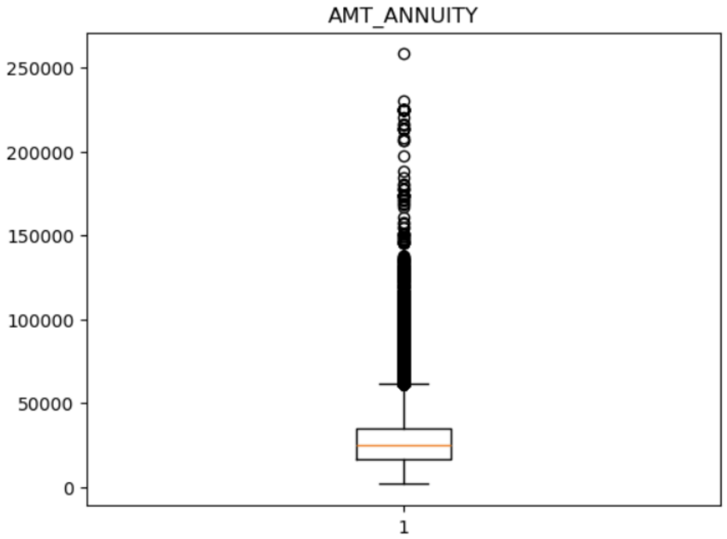
```
lowerwhisker is  -1.5
upperwhisker is  2.5
```

The values grater than 2.5 are consideresd to be outliers.



```
lowerwhisker is  -537975.0
upperwhisker is  1616625.0
```

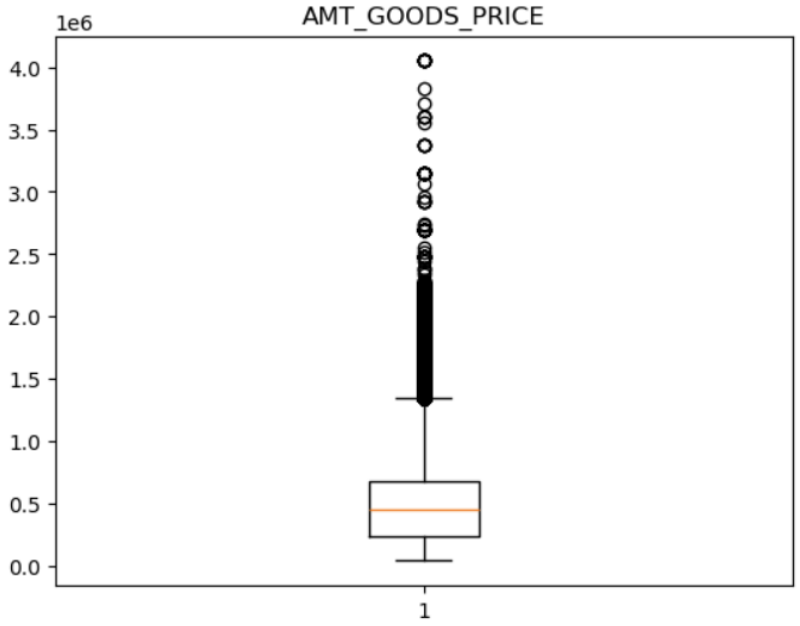
consider grater than 1616625.0 as outlier



```
lowerwhisker is  -10584.0
upperwhisker is  61704.0
```

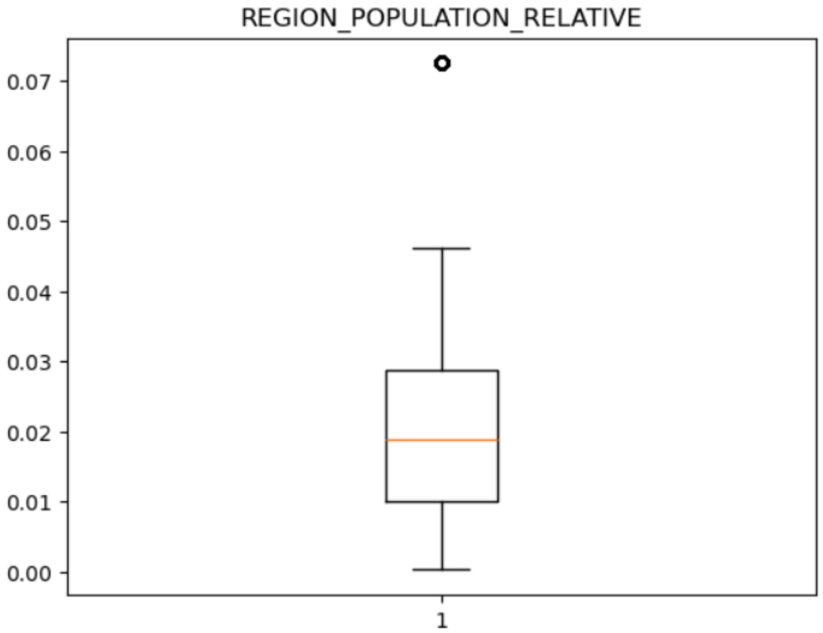
In this graph count grater than 61704.0 is consider as outlier.

Outliers



```
lowerwhisker is  -423000.0
upperwhisker is  1341000.0
```

In this graph count grater than 1341000.0 is consider as outlier

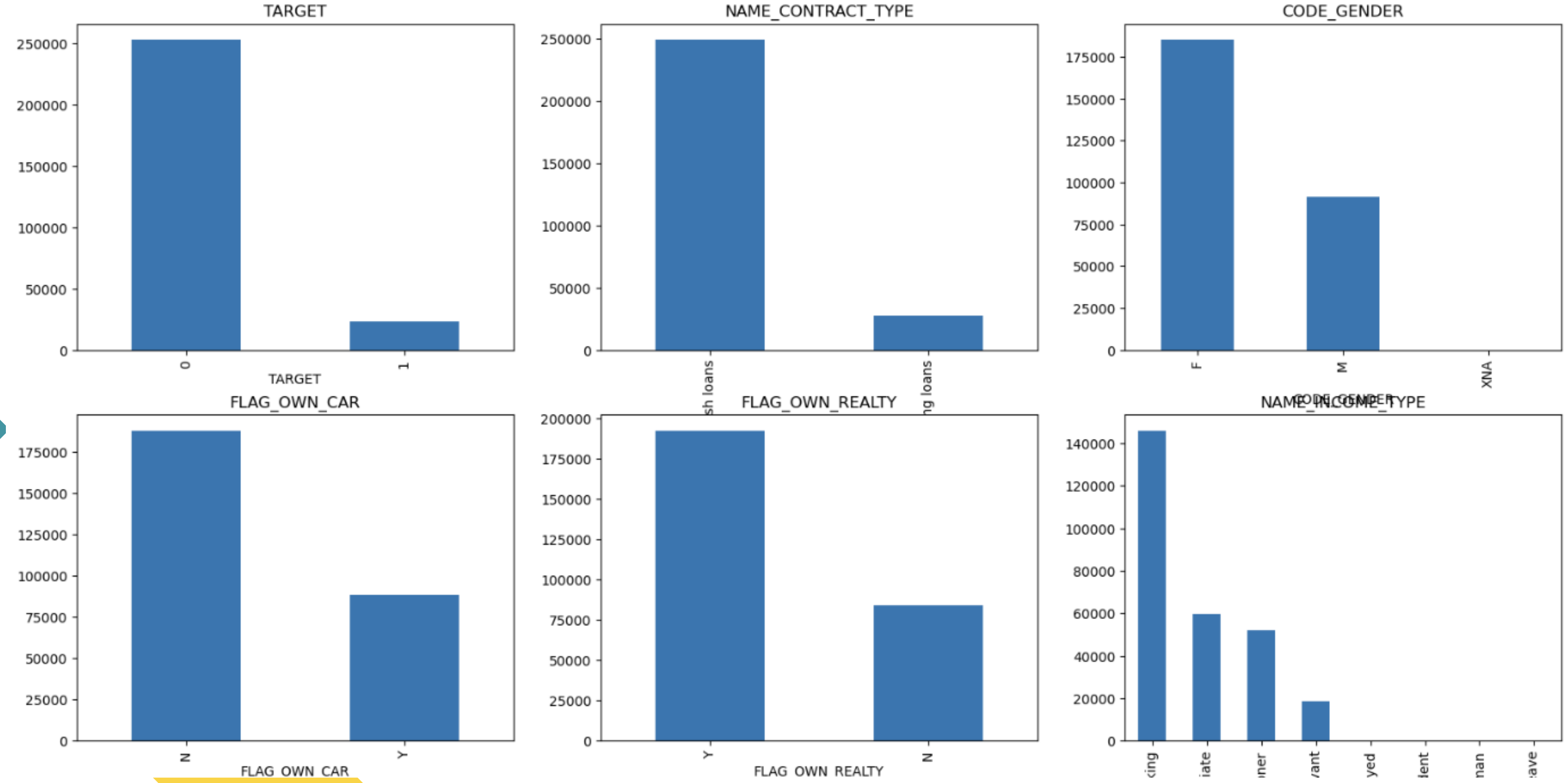


```
lowerwhisker is  -0.017979500000000002
upperwhisker is  0.056648500000000004
```

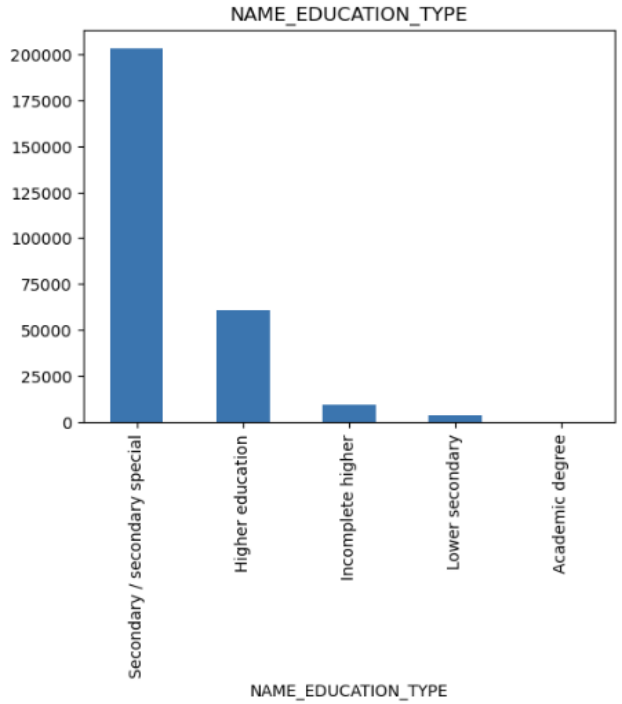
In this graph count grater than 0.05664 is consider as outlier

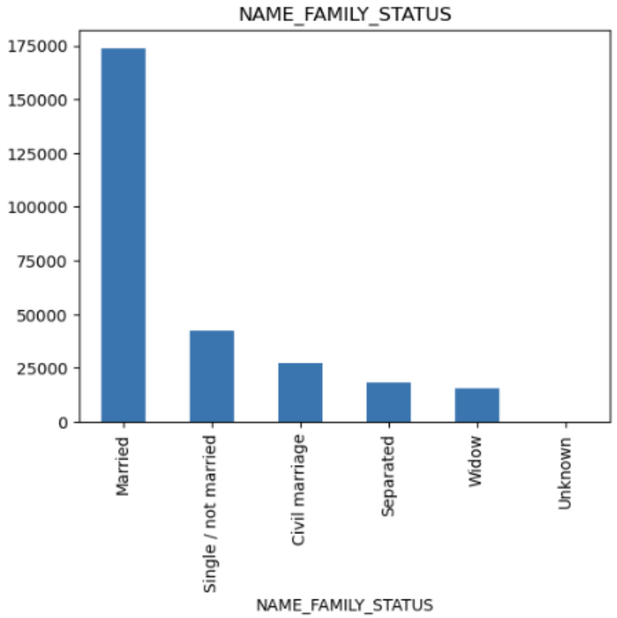
Data Imbalance

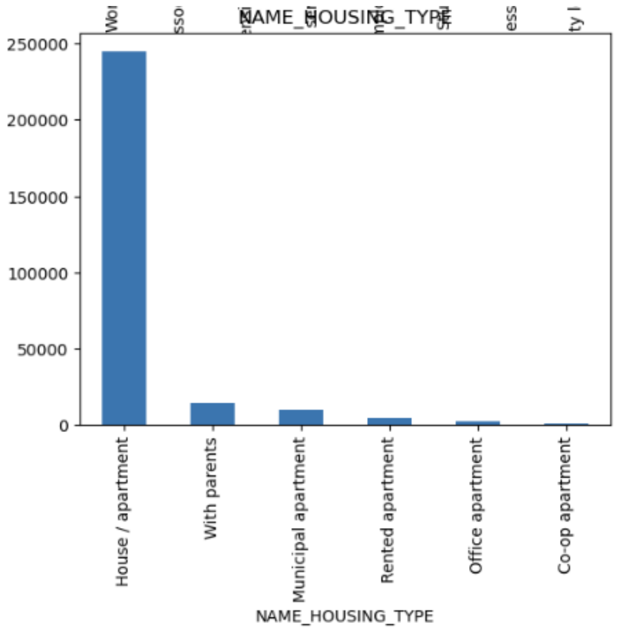
data imbalance refers to a disproportionate distribution of classes or categories in a dataset, which can bias analyses and predictive models.



Data Imbalance







TARGET - There are very few defaulters(1) compare to non defaulters(0)

NAME_CONTRACT_TYPE - There are very few Revolving loans than Cash loans

NAME_EDUCATION_TYPE - Most of the loans applied by Secondary/Secondary special educated people

NAME_FAMILY_STATUS - Most of the loans applied by Married people.

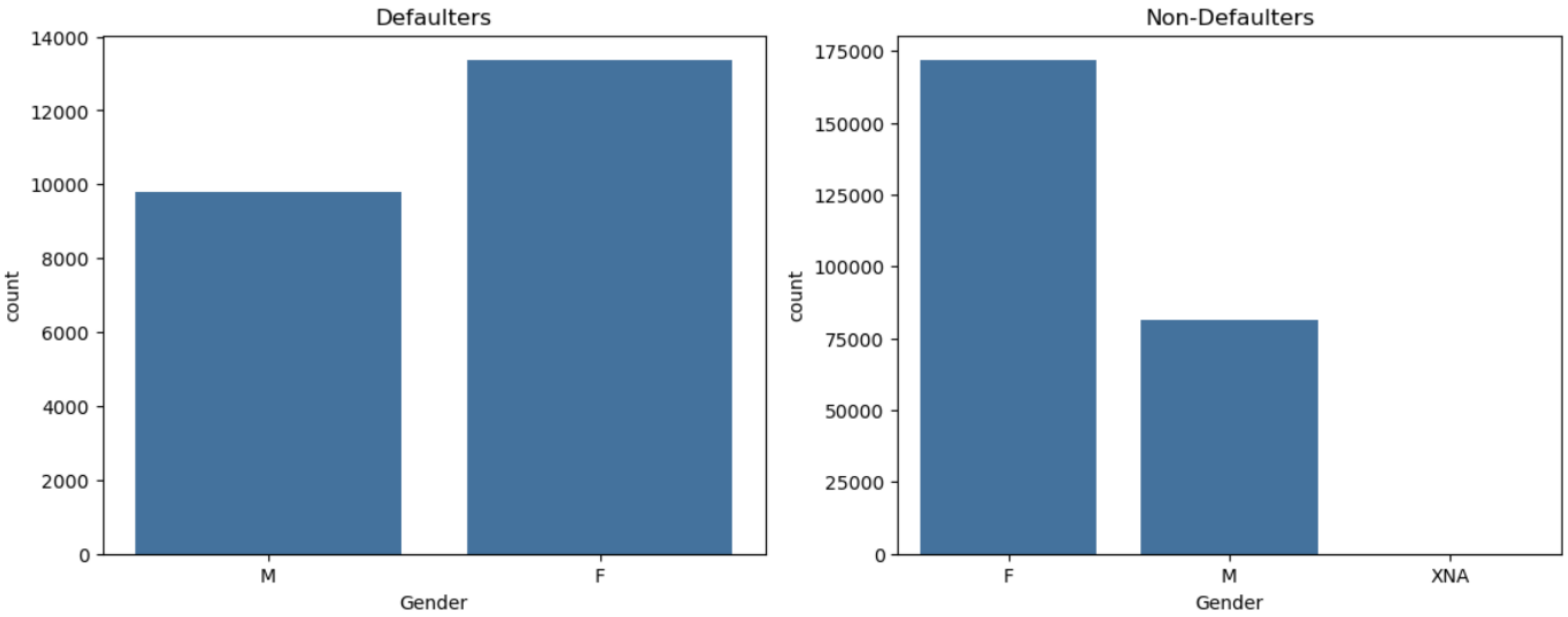
NAME_HOUSING_TYPE - Most of the application came from Home/apartment owner

Analysis

- **Univariate -> Univariate analysis examines a single variable**
- **Bivariate -> Bivariate analysis explores the relationship between two variables**
- **Multivariant -> Multivariate analysis investigates interactions among three or more variables**

Univariate Analysis on Application Data

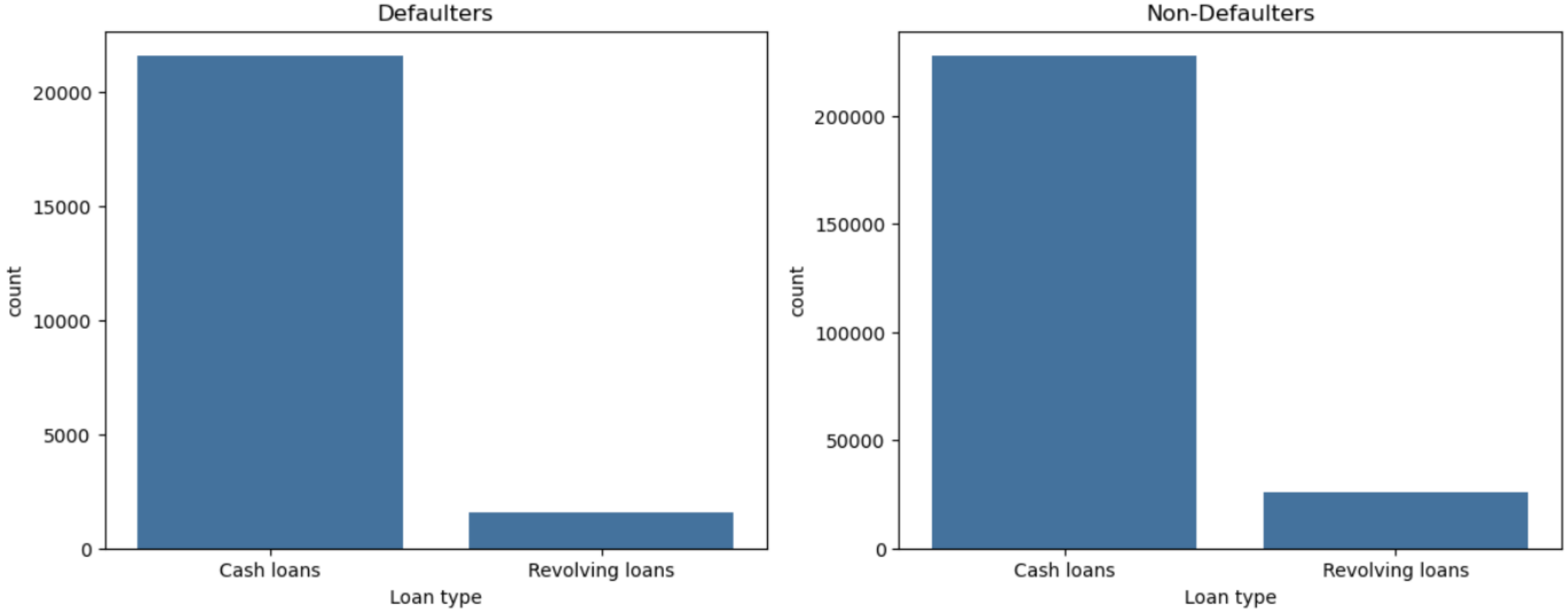
Analysis on basis on Gender -



- Defaluters - We can see that females are slightly more in number of defaulters than male.
- Non-defaluters - The same pattern continues for non-defaluters as well. The females are more in number here than male.

Univariate Analysis on Application Data

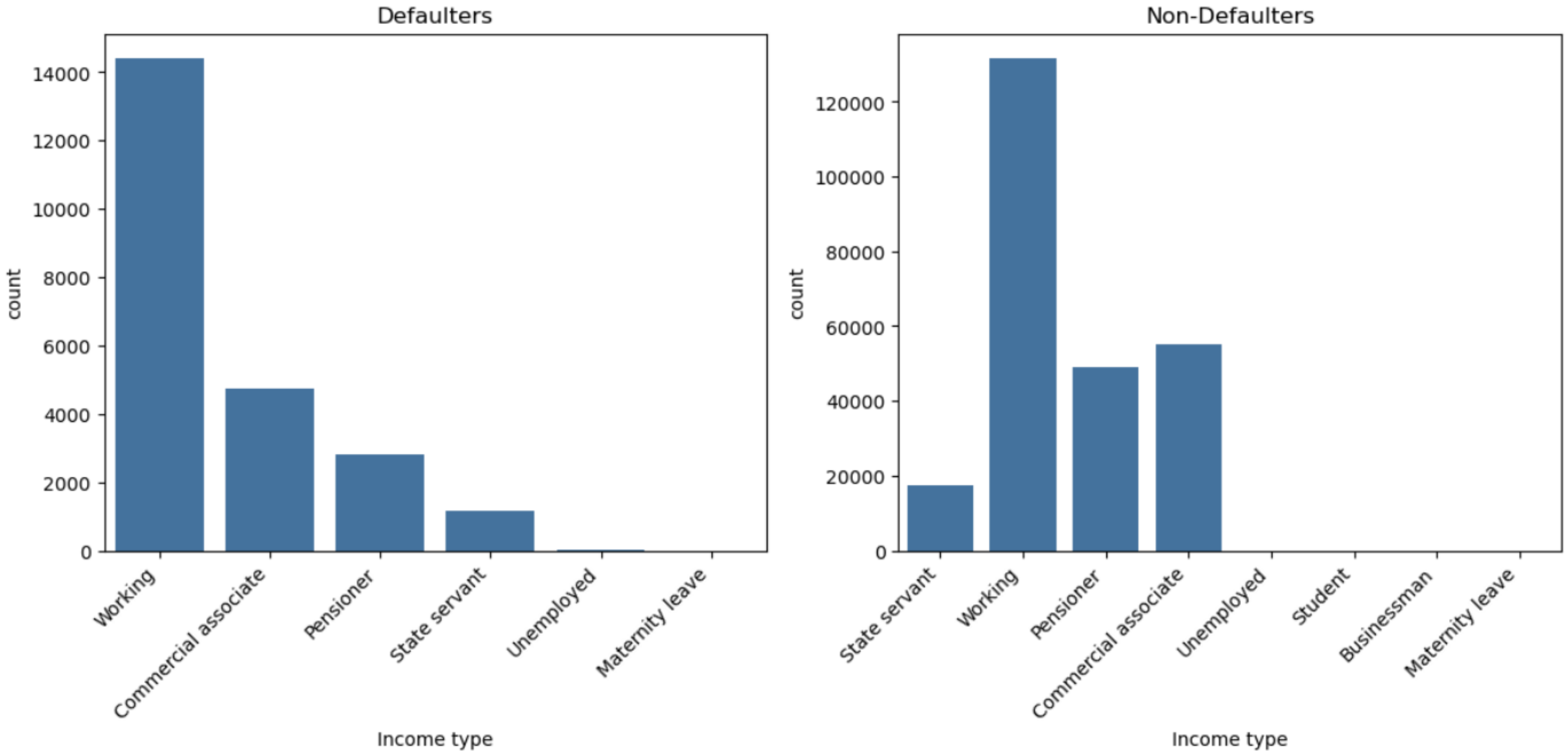
Analysis on basis on Loan Type



In this case Defaulter and non-defaulter both took cash loan as compare to revolving loans.

Univariate Analysis on Application Data

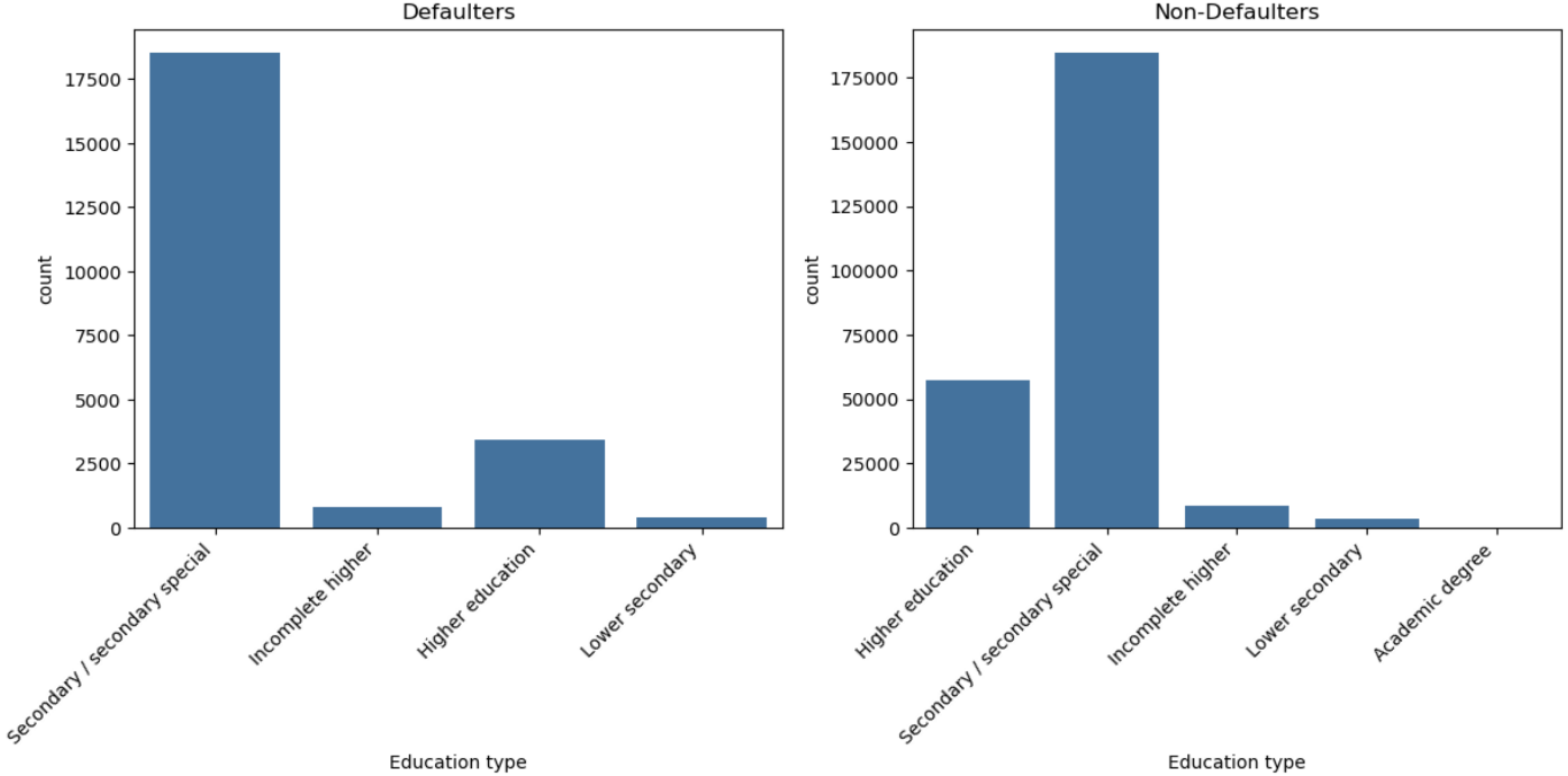
Analysis on basis on Income Type



- Defaulters - Working people are mostly defaulted as their numbers are high with compare to other pfrofessions.
- Non-defaulters - Similarly here also working people are more in number who are not defaulted.

Univariate Analysis on Application Data

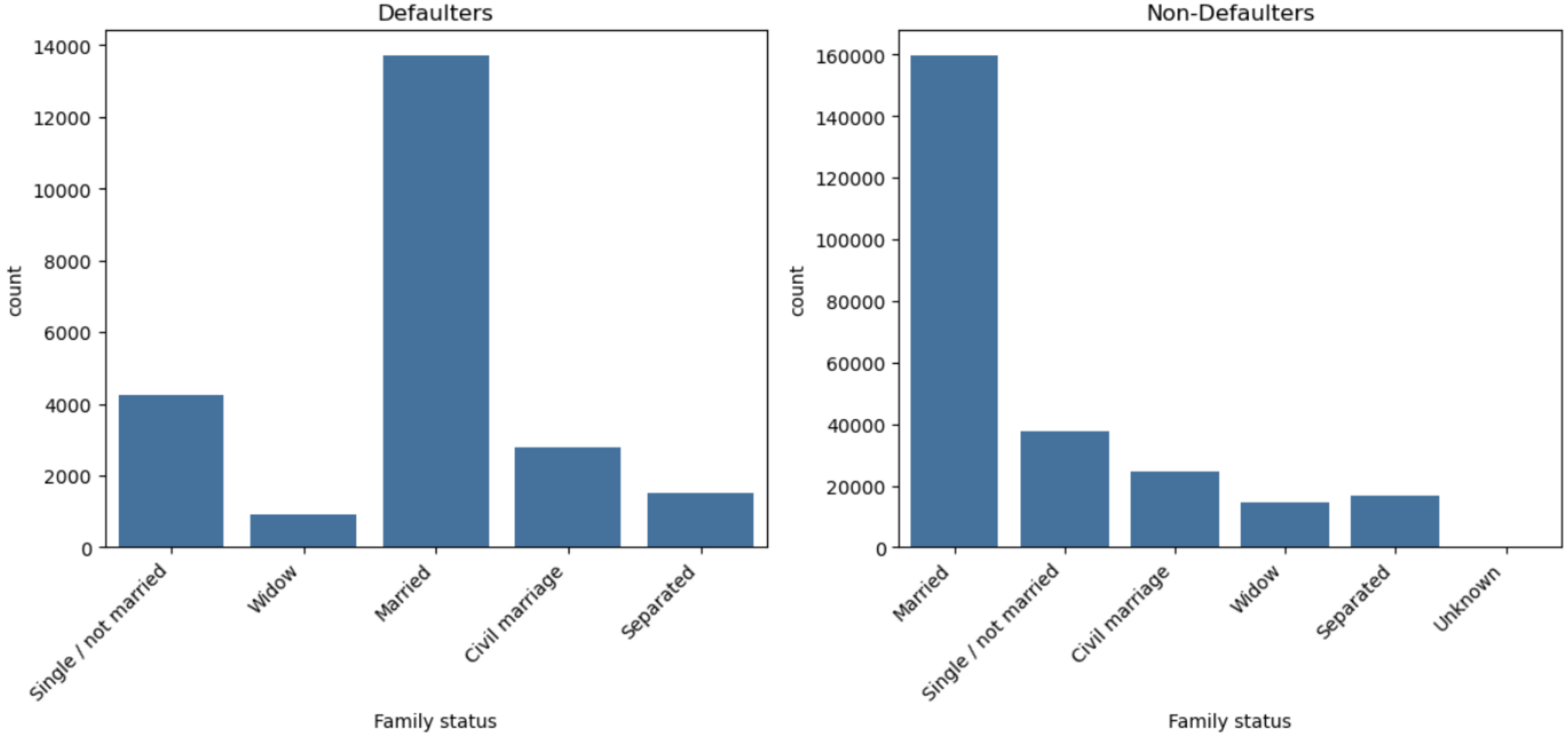
Analysis on basis on Education Type



- Defaulters - Education with Secondary/Secondary sepcial customers are more number in defaulters comapre with other level of eduacted poeple.
- Non defaulters - Here also Secondary/Secondary sepcial are more in numbers.

Univariate Analysis on Application Data

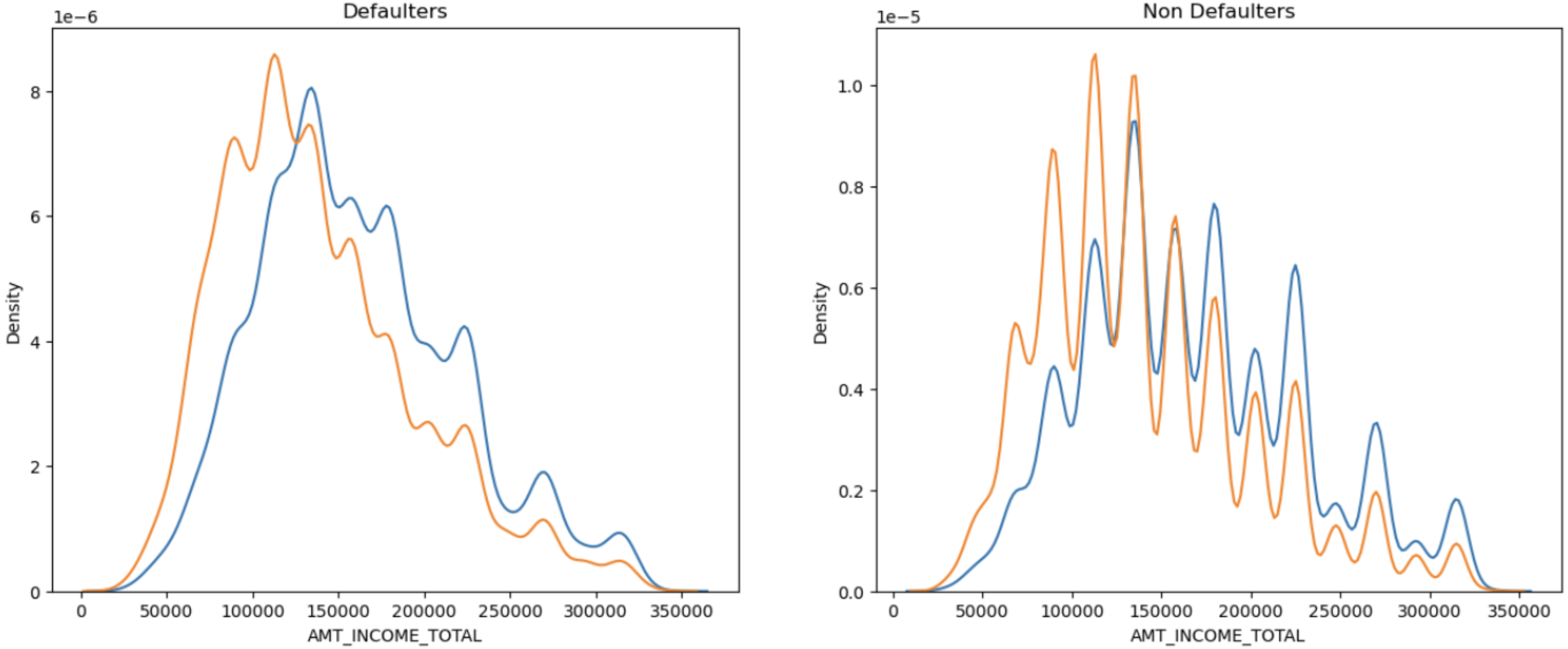
Analysis on basis of Family Status



For both the customers (defaulters and non-defaulters) married people are more in number comapred with single, separated, widow etc

Univariate Analysis on Application Data

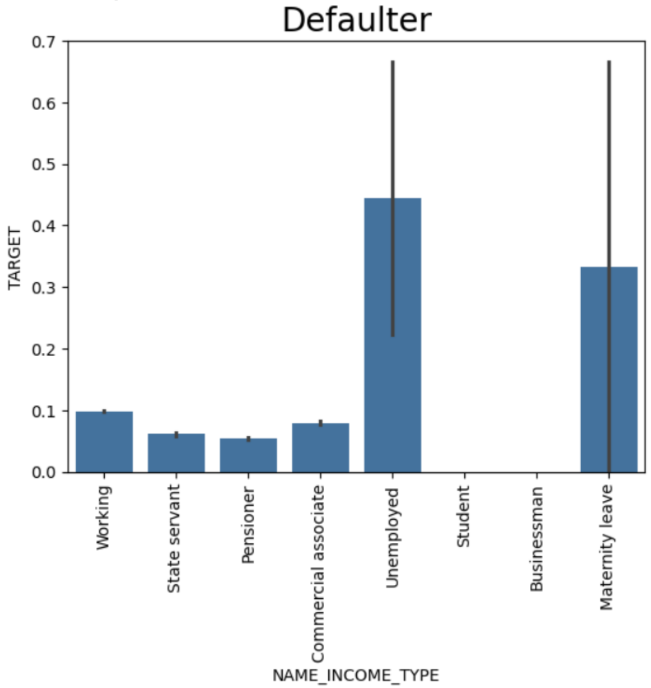
Analysis of basis of gender and their total income



- Defaulters - We can notice by looking at the pattern that for being a defaulter both the genders (male and female) are almost equal in all income levels. The spike of being defaulters is from 50000 to 200000.
- Non defaulters - Here we see an interesting pattern. Females are more non defaulter on the lower income level but lesser non defaluter in higher income level. The spike is more for both the genders from 75000 to 150000.

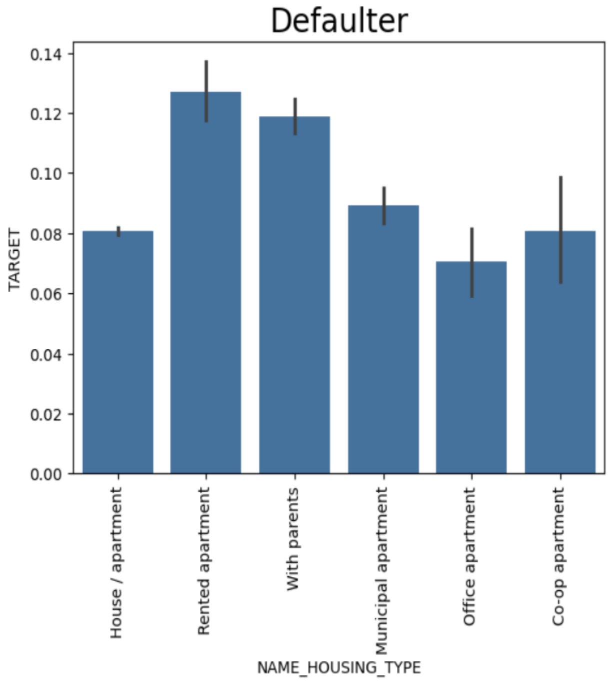
Bivariate Analysis on Application Data

Analysis on basis of Name_Income_Type Vs Target



Inferences from Barplot for Income Type: Highest percentage of defaulters are from unemployed applicants followed by maternity level applicants Less % of defaulters are from Pensioners, State Servants followed by working and Commercial associates(less than 10%).These categories can be considered for loans.

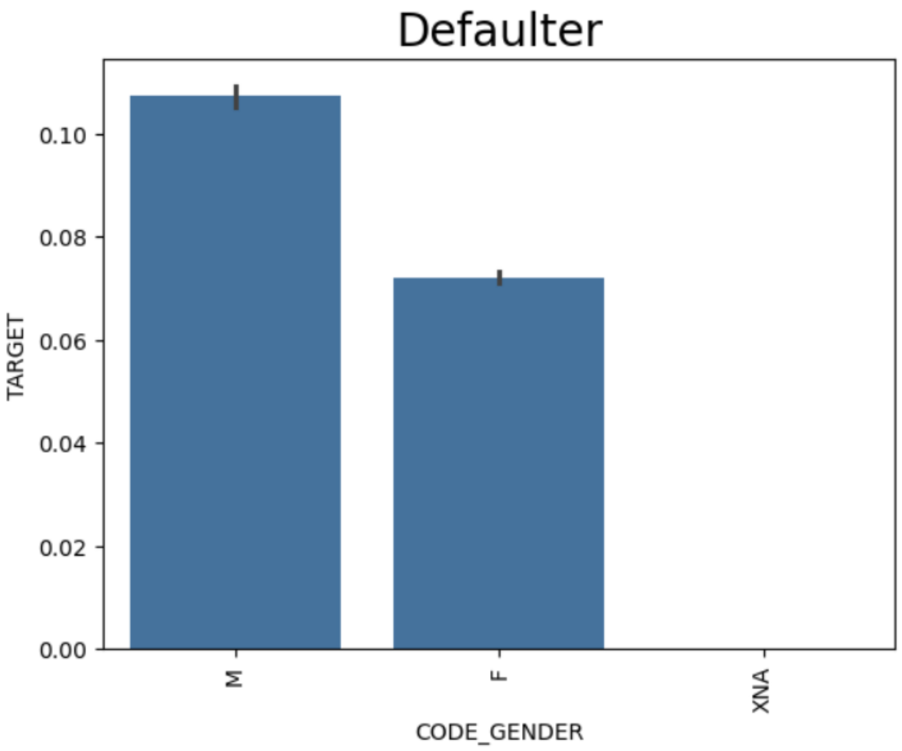
Analysis on basis of Name_Housing_Type Vs Target



Inferences from Barplot for Defaulters percentage in Housing Type: More defaulters are in Rented Apartment which is above 12% followed by Applicants living with parents. Less defaulters rate are in Office Apartment.

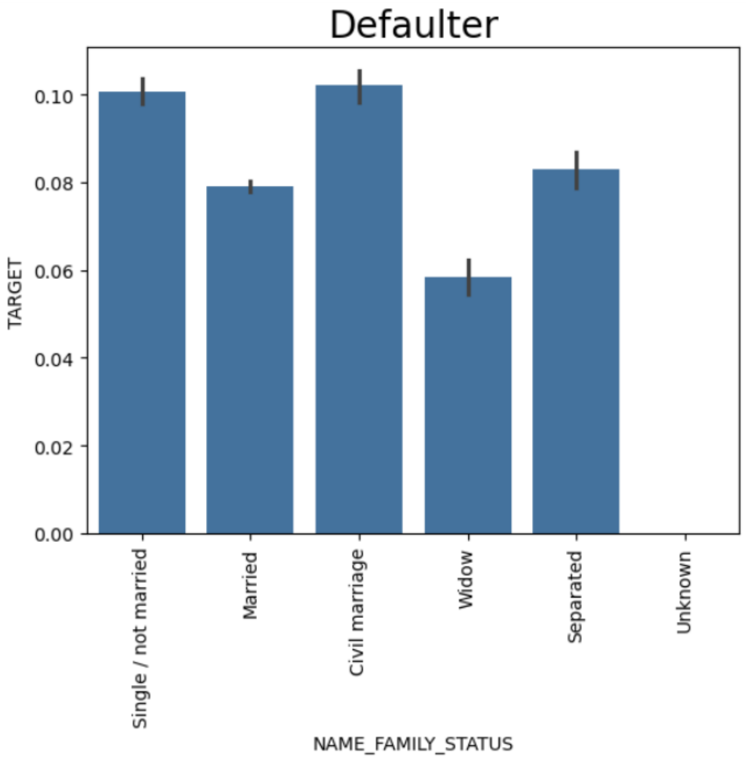
Bivariate Analysis on Application Data

Analysis on basis of Code_ Gender Vs Target



Inferences from Barplot for Gender: Defaulters% in Males are more compared to Female.

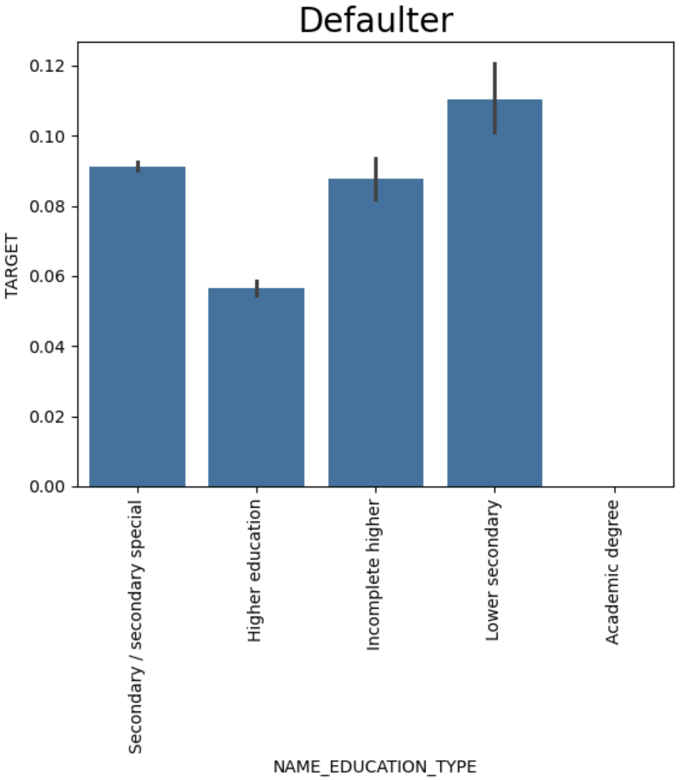
Analysis on basis of Name_Family_Status Vs Target



Inferences from Barplot for Family Status column: Civil Marriage are having highest % of defaulters,followed by Single/Not married.

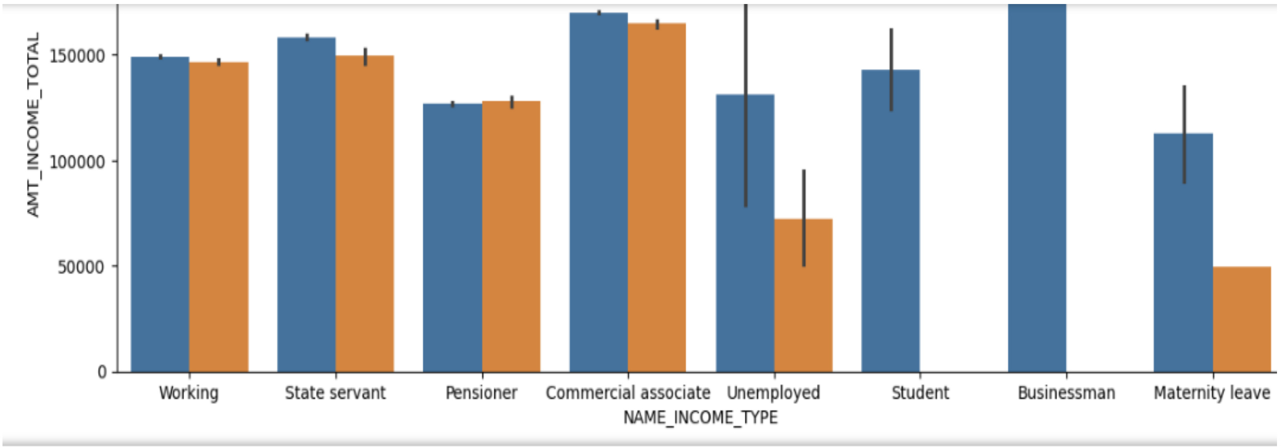
Bivariate Analysis on Application Data

Analysis on basis of Name_Education_Type Vs Target



Analysis from Barplot for Family Status: 11% of defaulters are in Lower Secondary and 9% defaulters in Secondary/secondary special education. Default percentage is less in Academic Degree applicants.

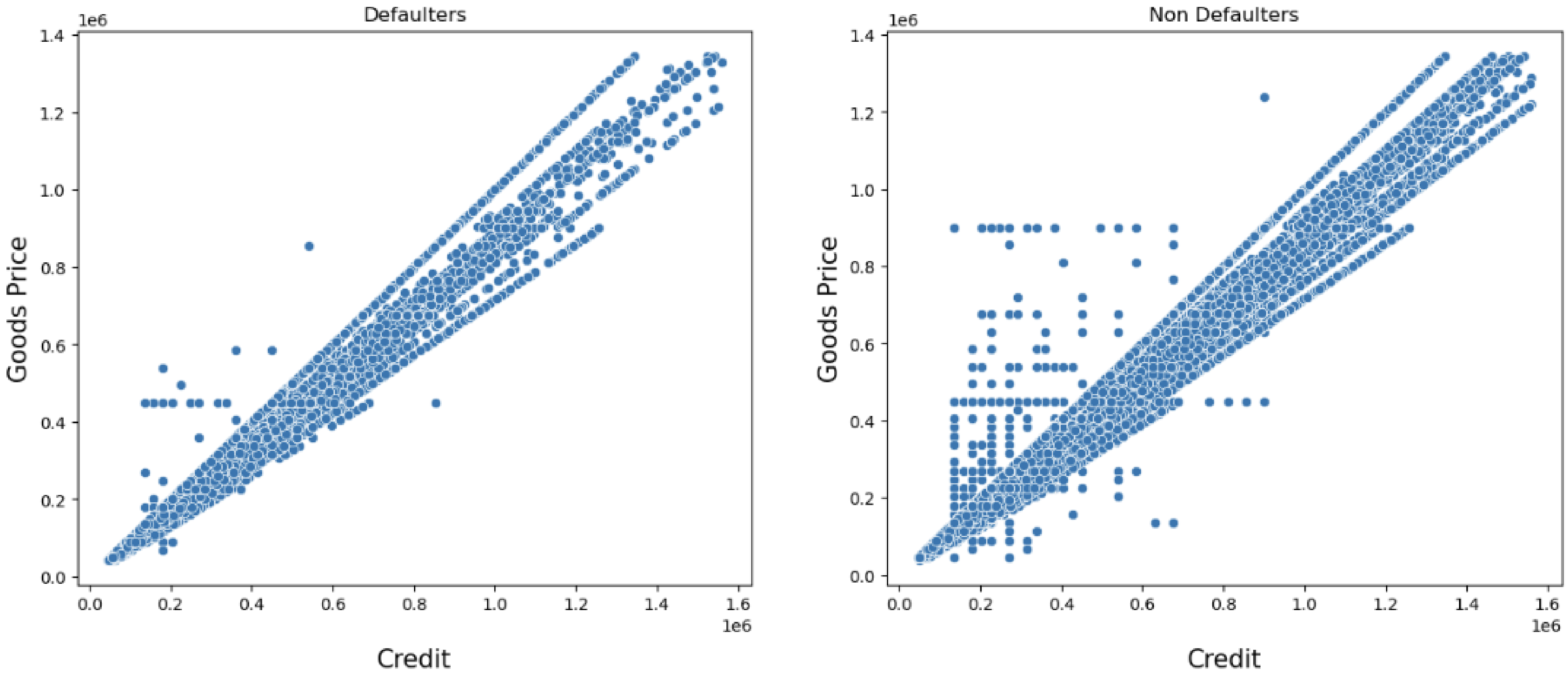
Analysis on basis of Name_Income_Type Vs AMT_Income_Total



No defaulters for Businessman and student having income. Bank can consider them for loan.

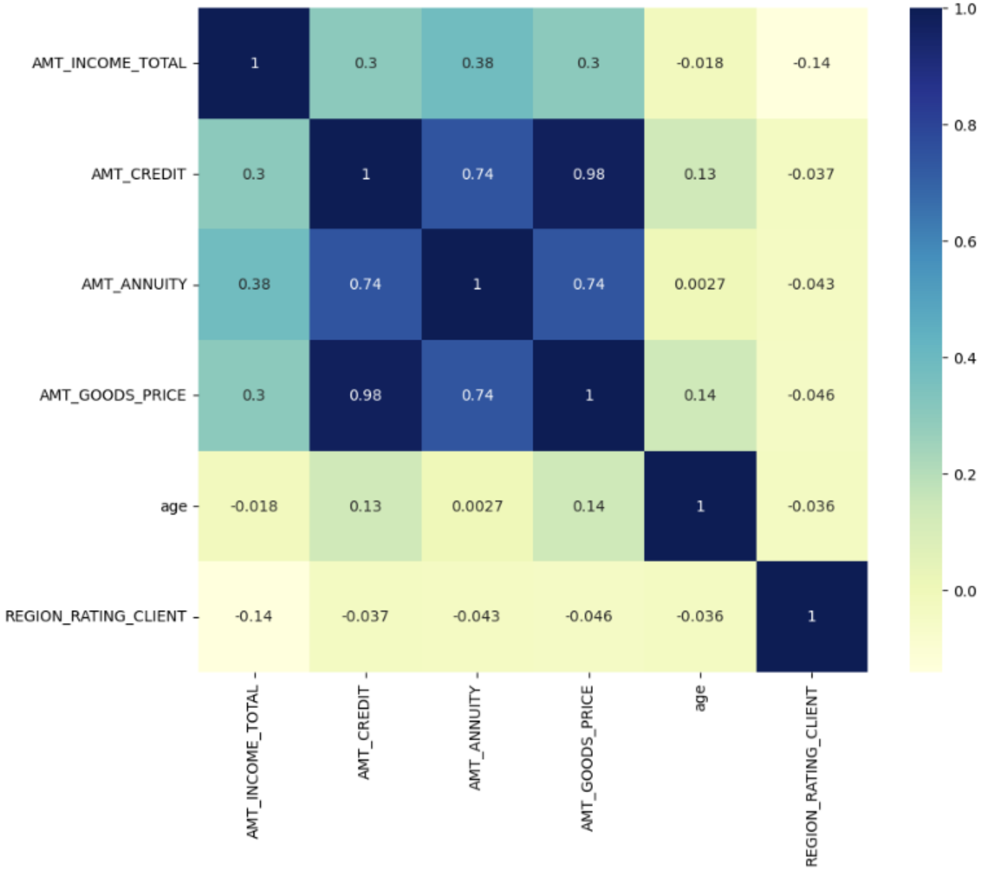
Bivariate Analysis on Application Data

Analysis AMT_Goods_Price and AMT_Credit using Scatter plot



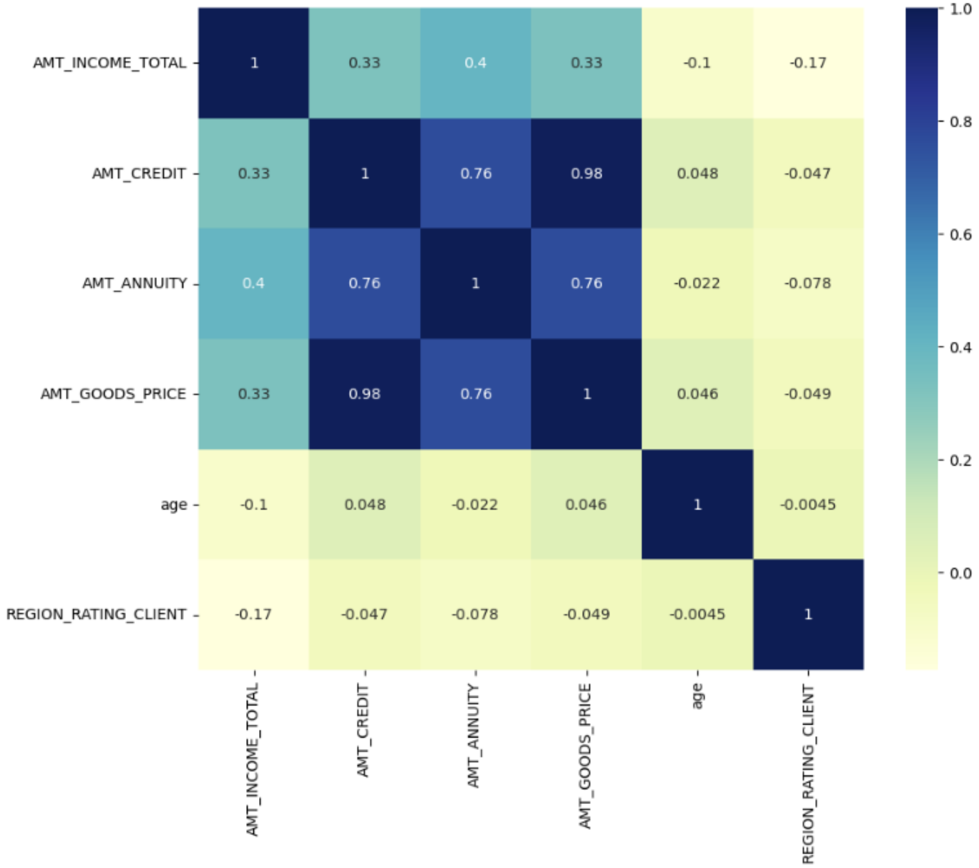
Analysis:AMT CREDIT and AMT GOODS PRICE are highly correlated,which means if increase in goods price,the credit increased directly.

Correlation - Defaulter & Non-Defaulters



Highly corelate columns for defaulters

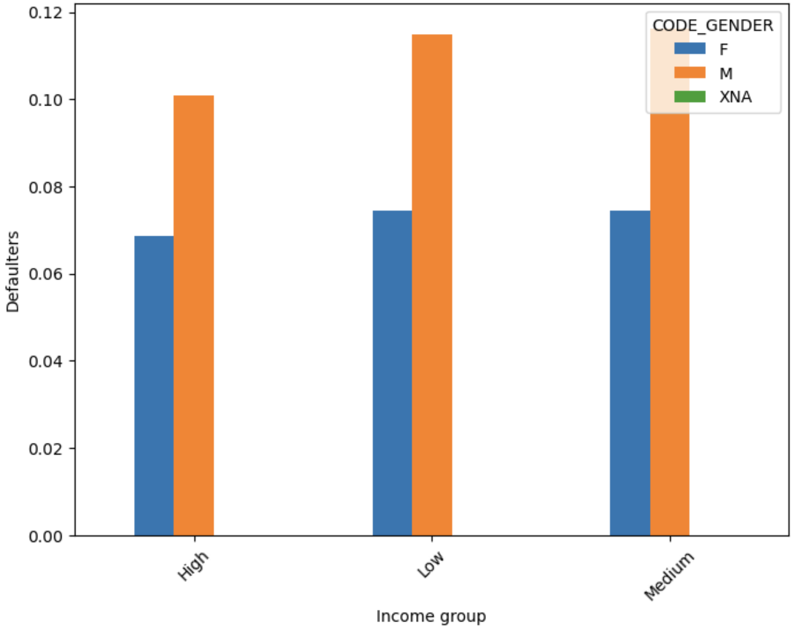
- AMT_CREDIT and AMT_ANNUITY (0.74)
- AMT_CREDIT and AMT_GOODS_PRICE (0.98)
- AMT_ANNUITY and AMT_GOODS_PRICE (0.74)



Highly corelate columns for non defaulters

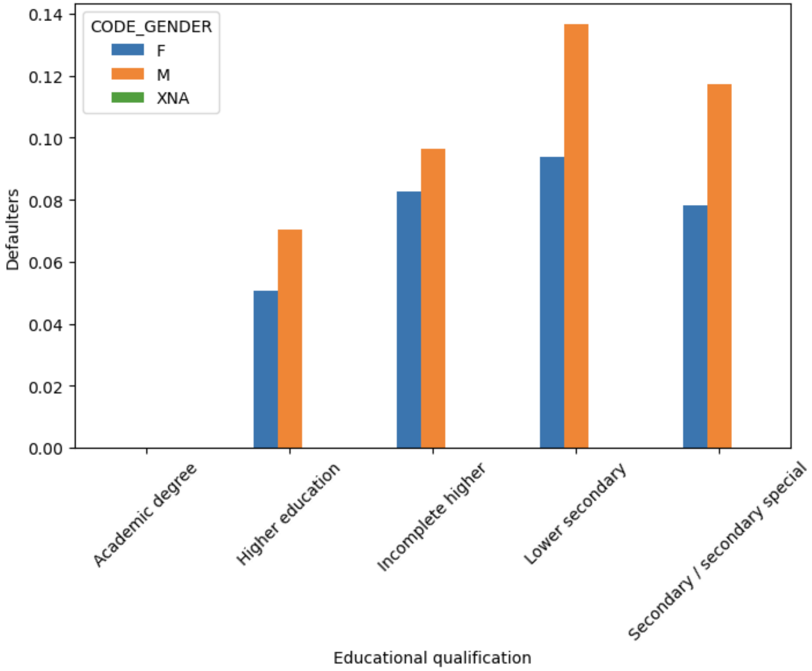
- AMT_CREDIT and AMT_ANNUITY (0.76)
- AMT_CREDIT and AMT_GOODS_PRICE (0.98)
- AMT_ANNUITY and AMT_GOODS_PRICE (0.76)

Analysis of Income group Vs Gender



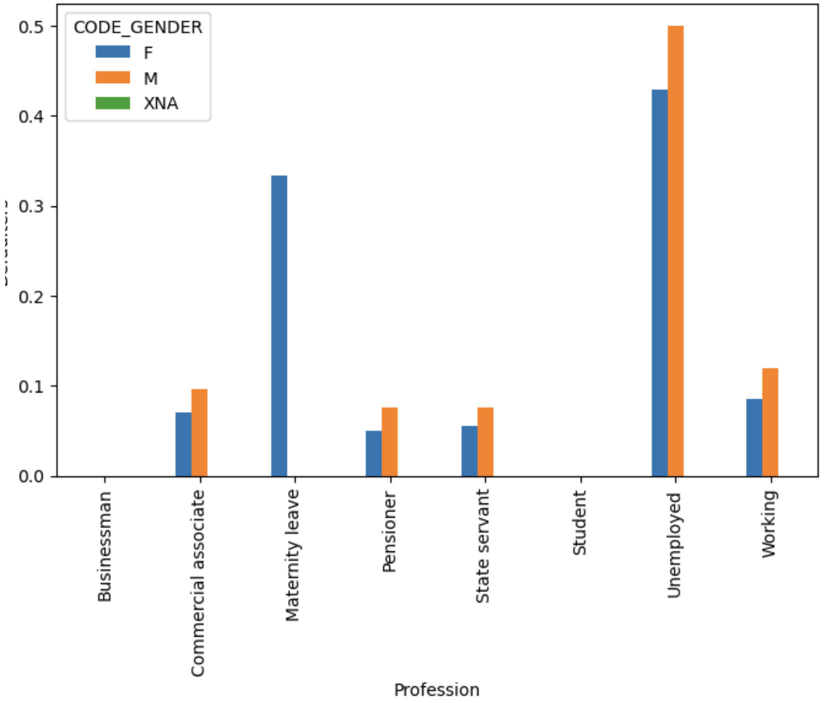
We can see that Males are more likely defaulted than Females accross all income groups.

Analysis of Education_Type Vs Gender



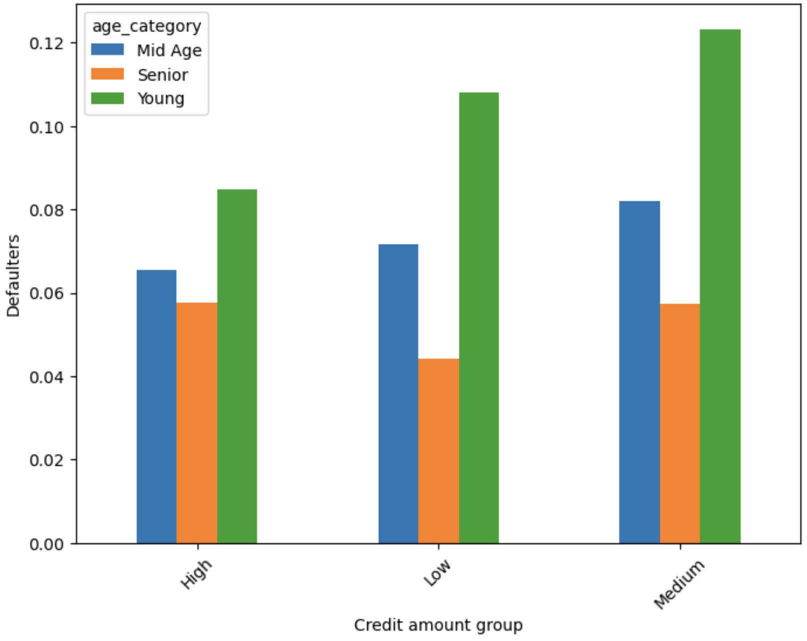
- Lower secondary educated clients are more defaulted followed by Secondary and Incomplete higher educated clients.
- The Higher educated people are less defaulted.
- Accross all educated level Females are less defaulted than male.

Analysis of Profession Vs Gender



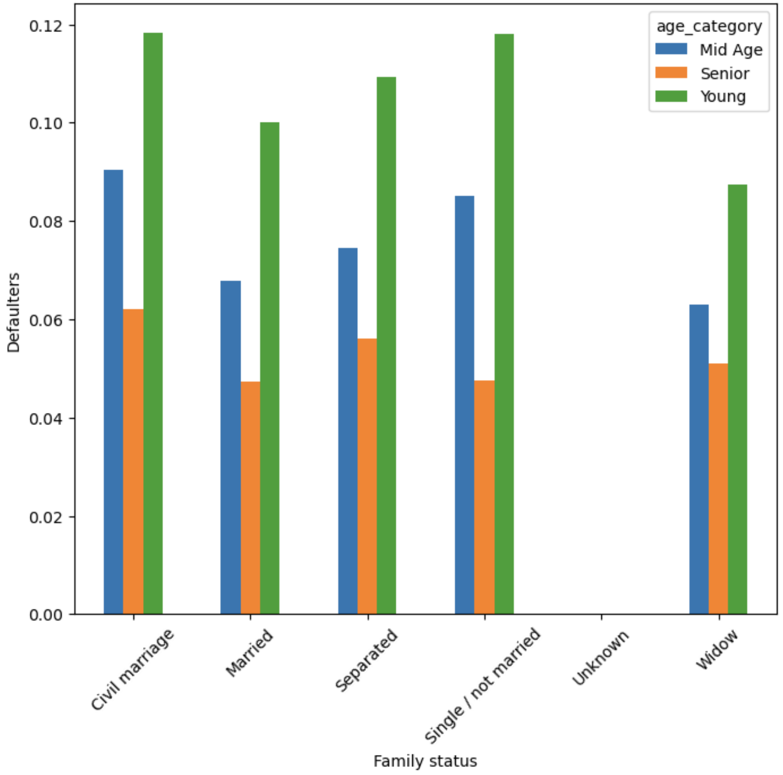
- No surprise the unemployed clients are more defaulted.
- Clients with maternity leave are expected to be defaulted more.
- The default rate is lesser in all other professions.
- Males are more defaulted with their respective professions compared to females.

Analysis of Credit amount group Vs Age group



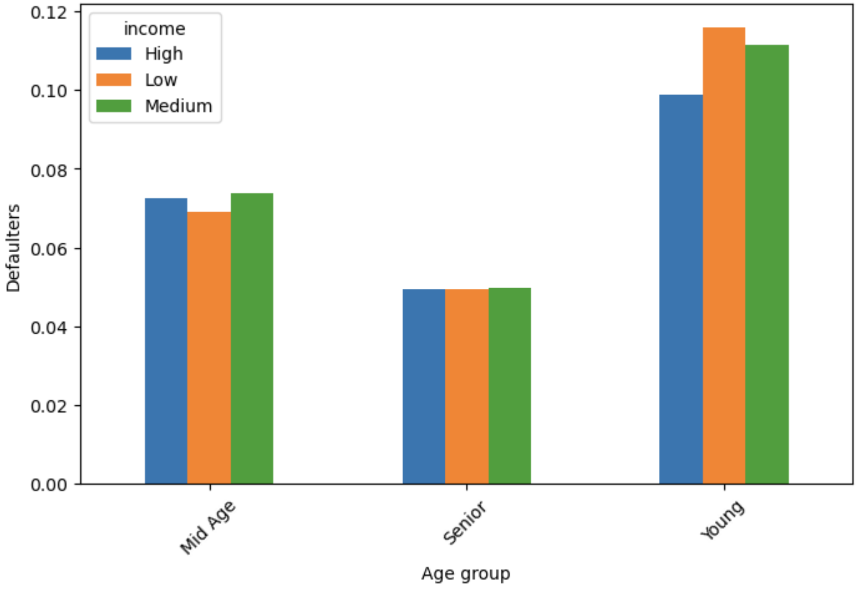
- Young clients with medium and low credit amount group are highly defaulted.
- Senior citizens across all credit amount groups are less likely defaulted.

Analysis of Family Status Vs Age Group



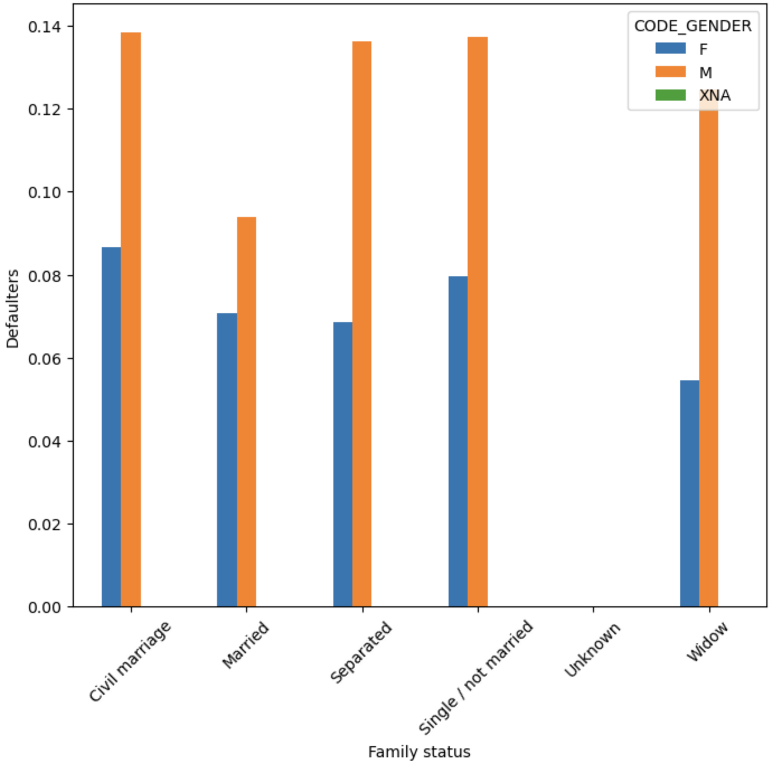
Across all family status the Young clients are more defaulted and Senior citizen are less.

Analysis of Income group Vs Age Group



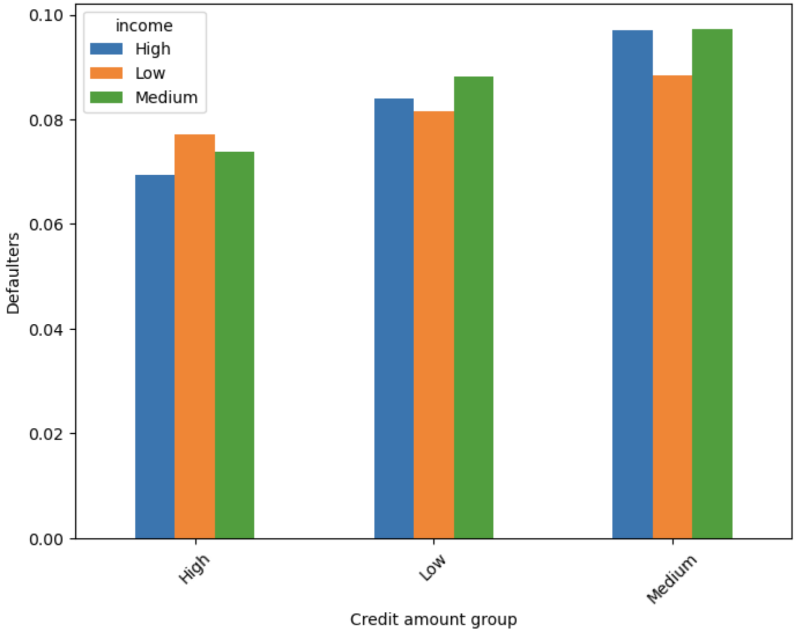
- Young clients are more defaulted than Mid age and senior.
- Young low income people are more defaulted.
- For Mid age and senior people the default rate is almost same in all income group.

Analysis of Family status Vs gender



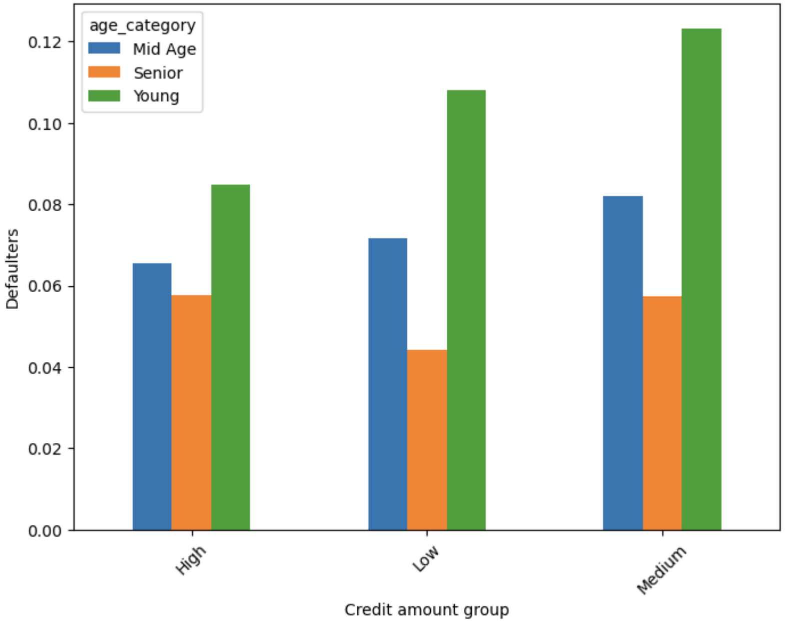
Across all family status the Male clients are more defaulted than Female.

Analysis of Credit_amount Vs Income group



- Medium credit amount group are highly defaulted in all income groups.
- High credit amount groups are less likely to default in all income groups.

Analysis of Credit_Amount Vs Age Group

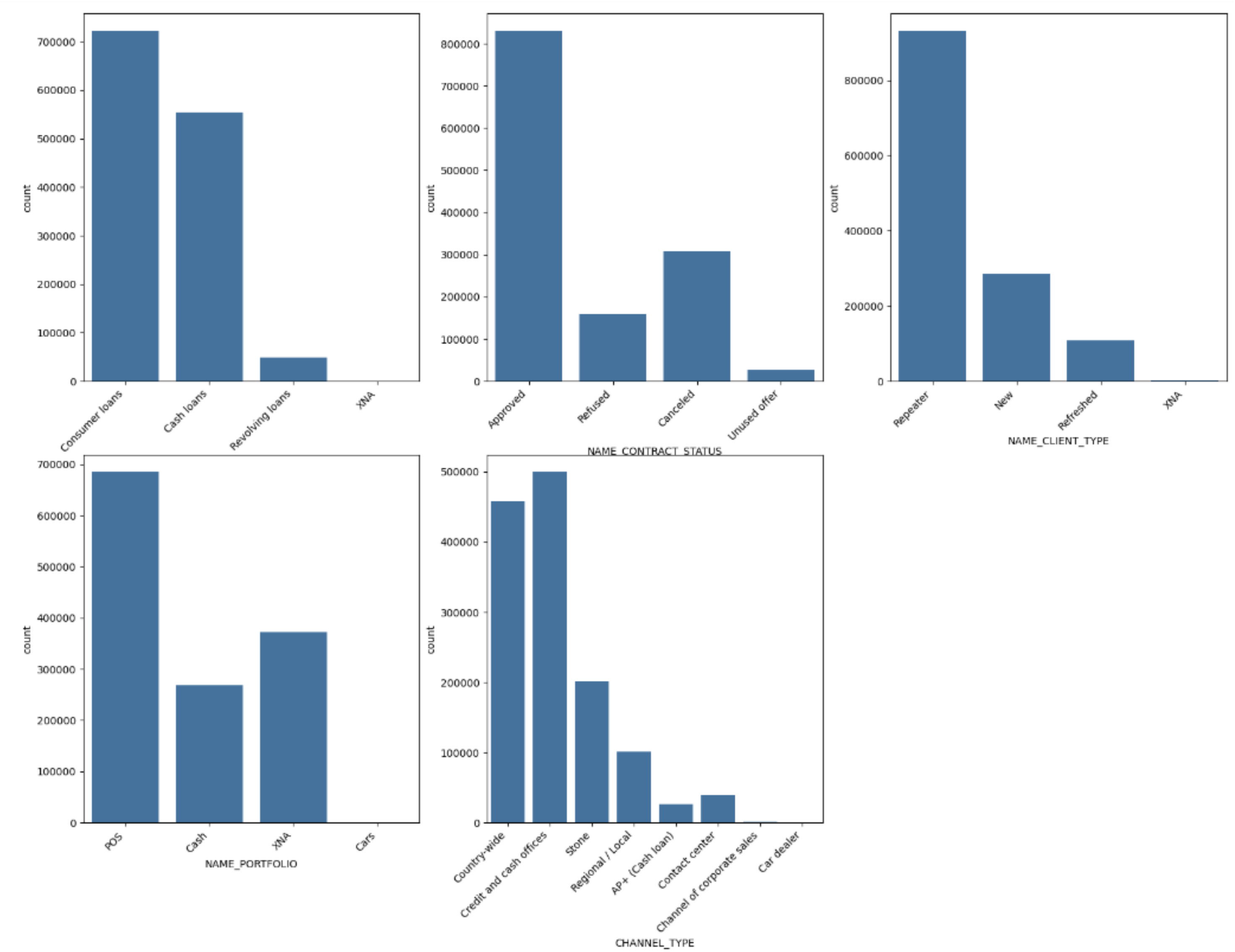


- Young clients with medium and low credit amount group are highly defaulted.
- Senior citizens across all credit amount groups are less likely defaulted.

Analysis of Previous Data Set

Univariate Analysis

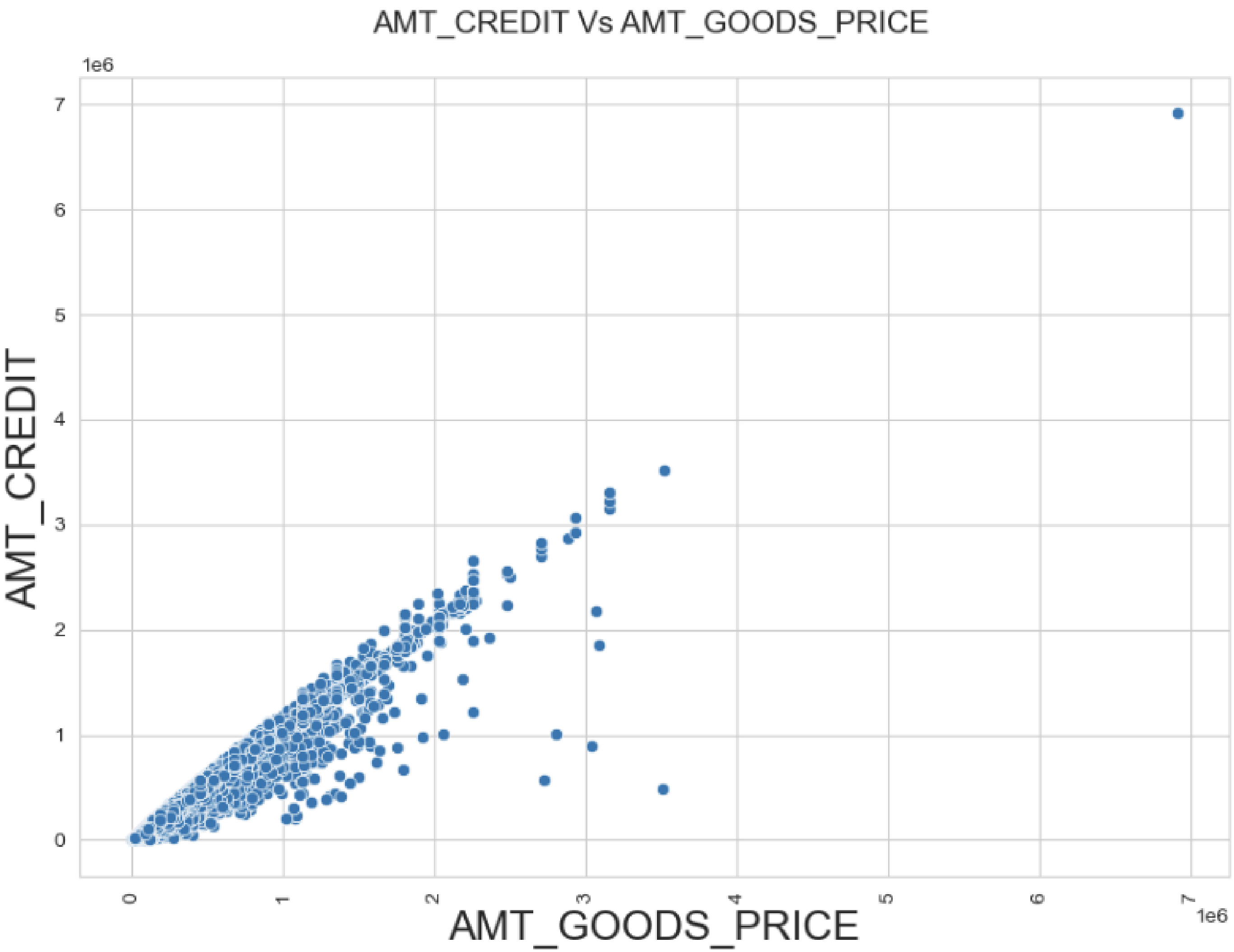
NAME_CONTRACT_TYPE - There are very few Revolving Loans
NAME_CONTRACT_STATUS - There are very few Refused loans, Almost negligible Cancelled loans.
NAME_CLIENT_TYPE - There are very few New applicant, Even fewer Refreshed applicants.
NAME_PORTFOLIO - Very few application for Cards and Cars
CHANNEL_TYPE - Except Country-Wide, Credit and Cash offices and Stone all other channels are very few in number.



Analysis of Previous Data Set

Bivariate Analysis

Analysis of AMT_Credit Vs AMT_Goods_Price



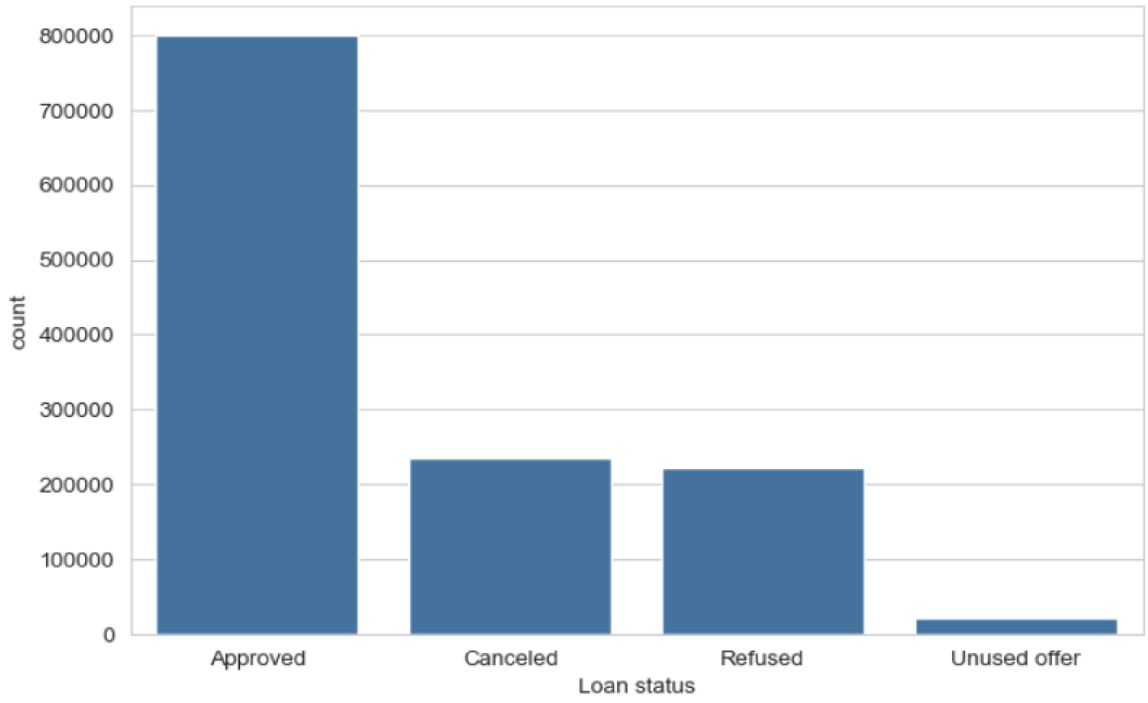
Merged both Application and Previous Data sets

Two datasets are merged to combine complementary information, enabling comprehensive analysis and uncovering relationships across datasets.

Here I am doing left join on the same column 'SK_ID_CURR', which merges two datasets by including all records from the left dataset and matching records from the right dataset, with unmatched entries filled as nulls.

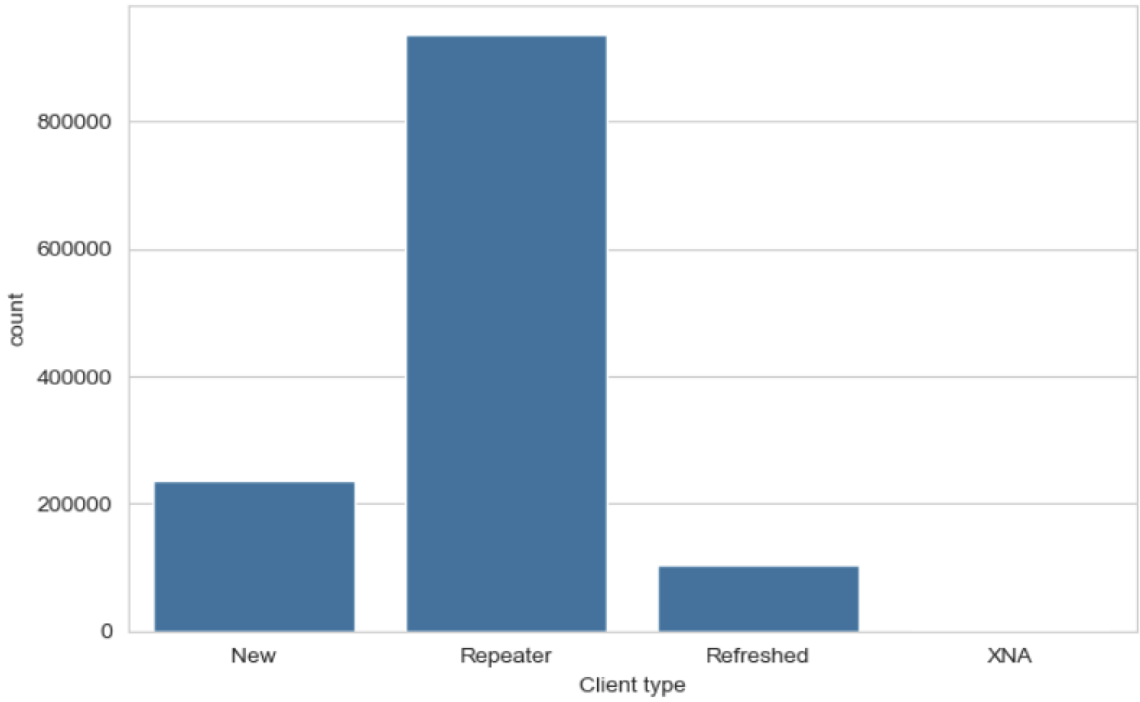
Univariant Analysis

Analysis of Name_Contract_Status



There are huge number of Approved loan than Refused and canceled. Hardly, there are any unused offer loan.

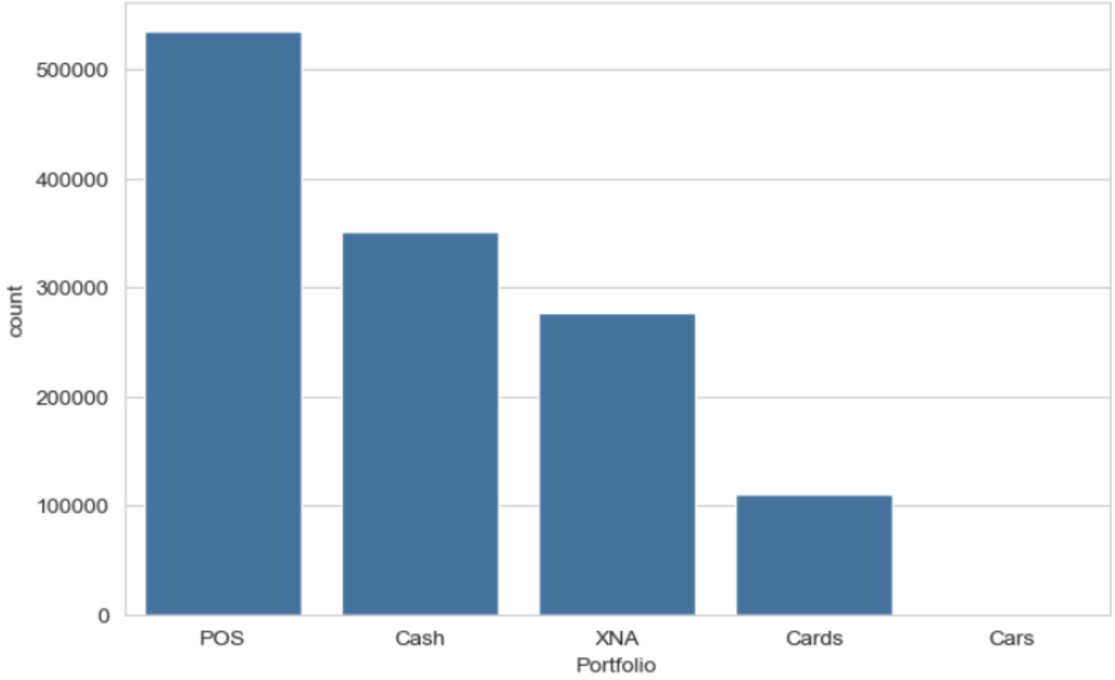
Analysis of Name_Client_Type



Mostly the applicants were Repeater

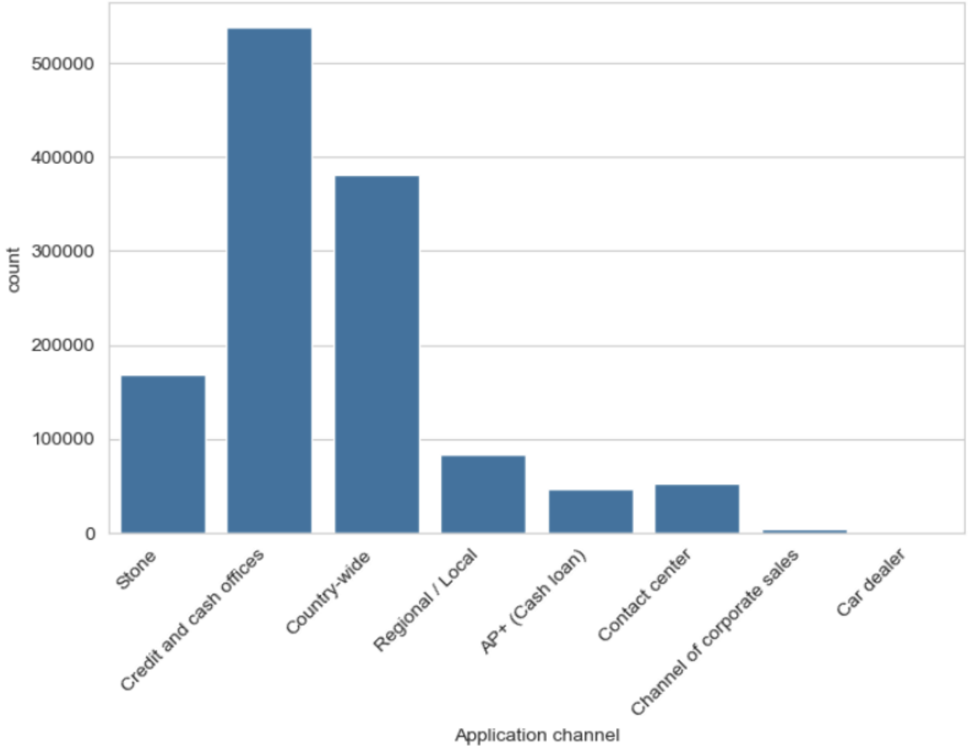
Univariant Analysis

Analysis of Name_Portfolio



The highest number of the previous applications was for POS. Applications for Cash also has good number.

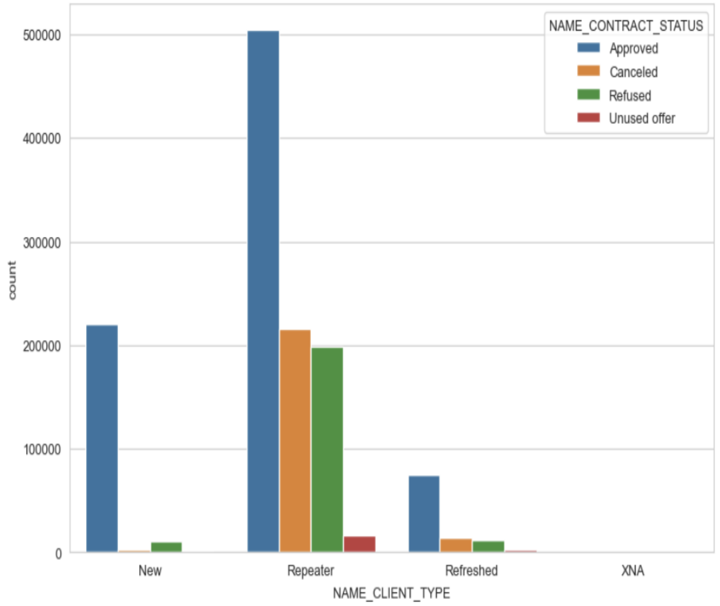
Analysis of Channel_type



We see that Credit and Cash offices was heavily used for previous applications followed by Country-wide, Stone. Rest other channels are hardly used.

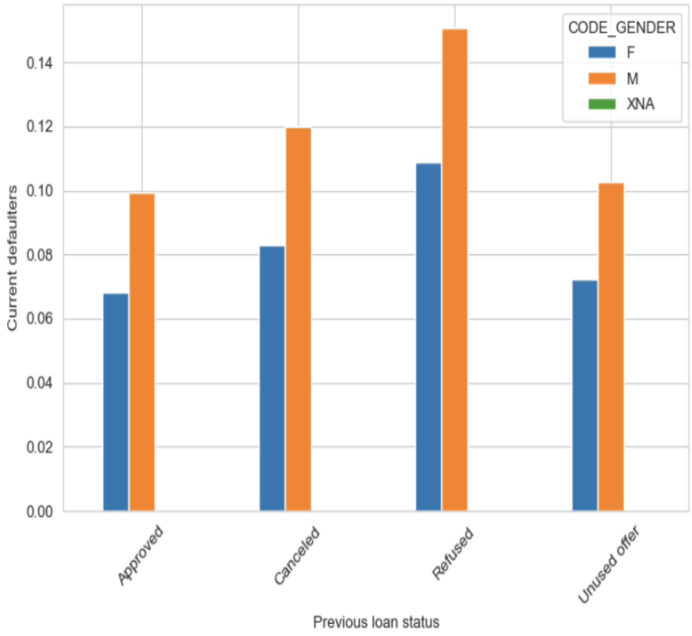
Bivariate Analysis

Analysis of Name_Contract_Status and Name_Client_Type



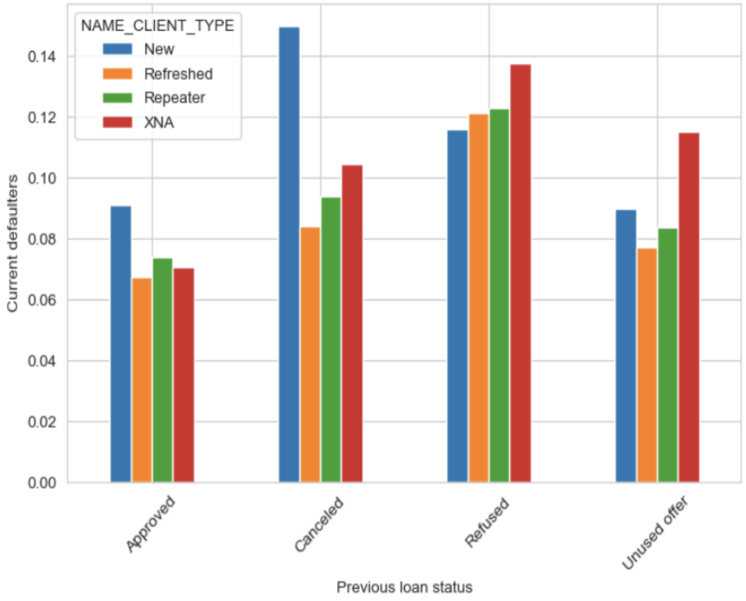
We see that the Repeater clients have more approved loans than New and Refreshed clients.

Analysis of Name_Contract_Status, Code_Gender and Target



We see that previously Refused client is more defaulted than previously Approved clients. Also, in all the cases the Males are more defaulted than Females.

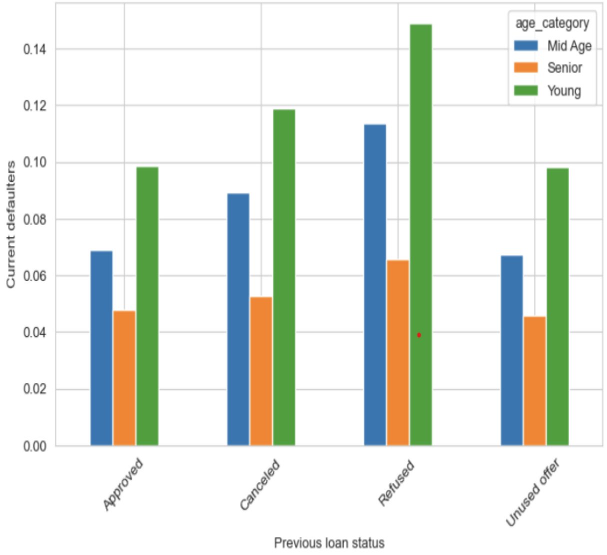
Analysis of Name_Contract_Status, Name_Client_Type and Target



- We can see that the Defaulters are more for previously Unused offers loan status clients, who were New.
- For previously Approved status the New clients were more defaulted followed by Repeater.
- For previously Refused applicants the Defaulters are more Refreshed clients.
- For previously Canceled applicants the Defaulters are more New clients.

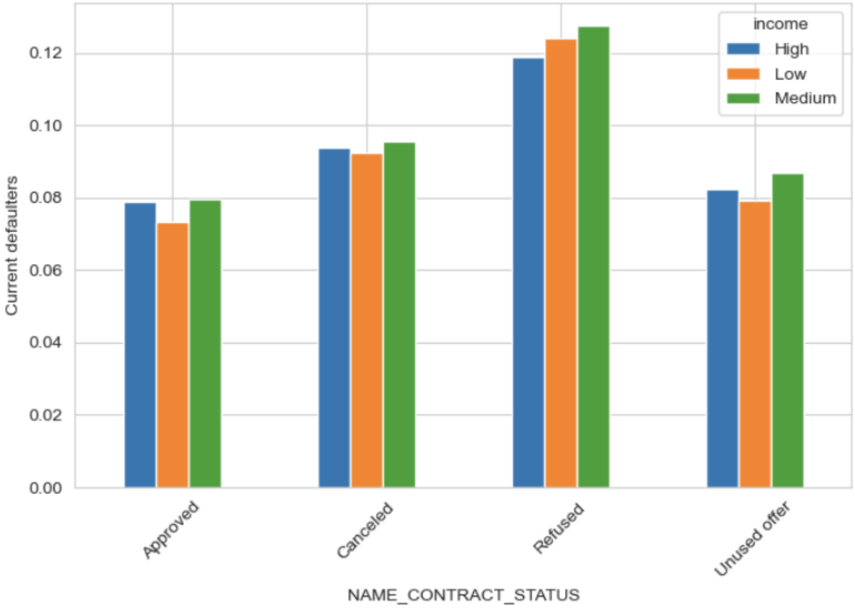
Bivariate Analysis

Analysis of Age_Category, Target and Name_Contract_Status



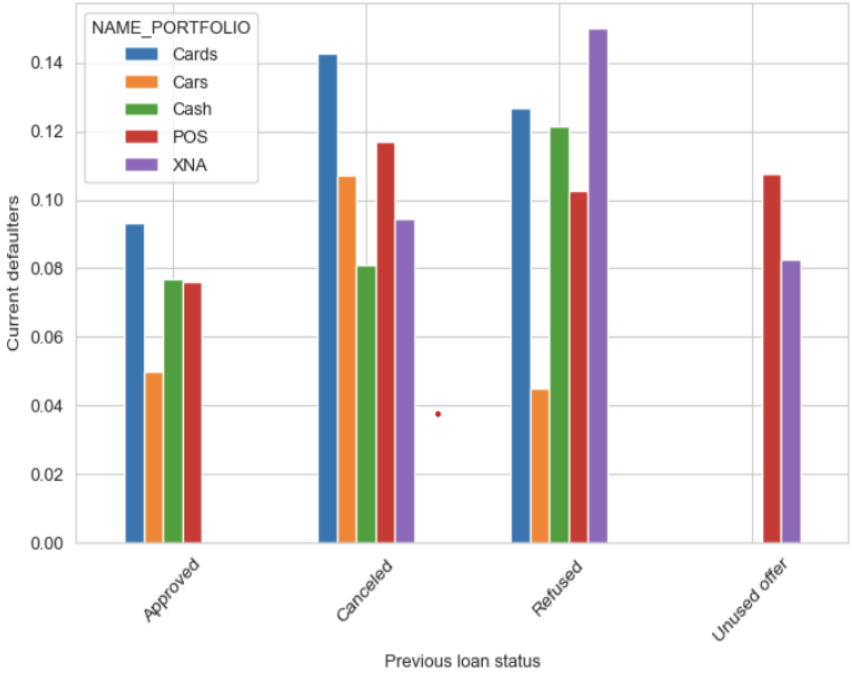
- For all the previous status Young applicants are more defaulted, Senior applicants are less defaulted compared to others.

Analysis of Name_Contract_Status, Income and Target



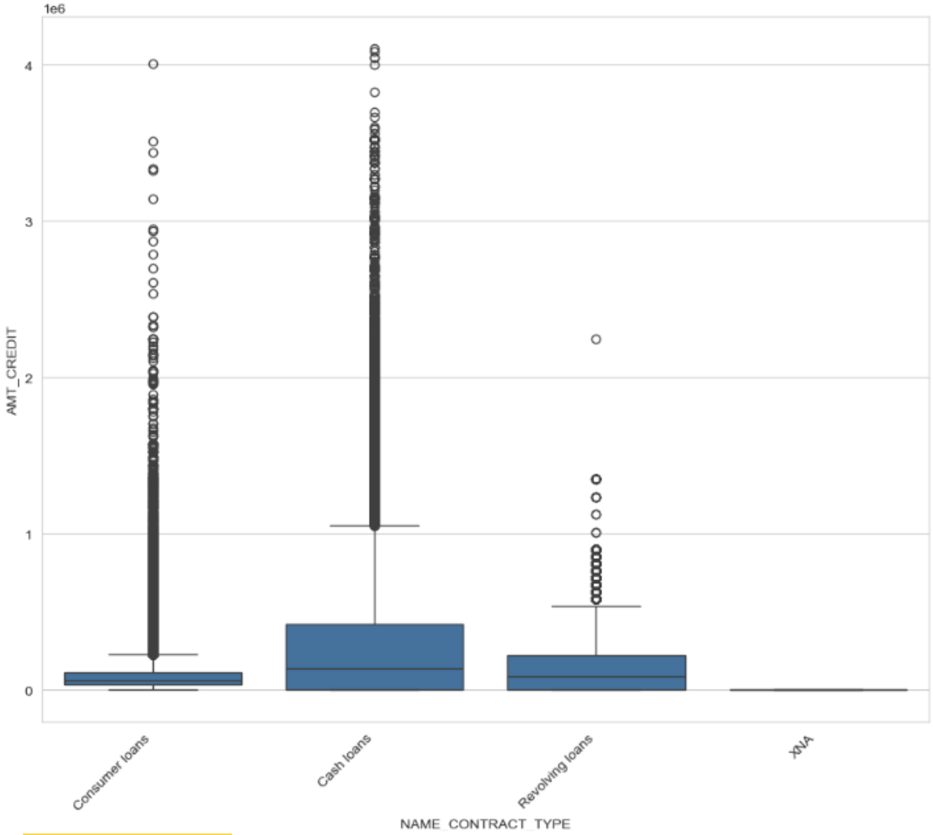
- For previously Unused offer the Medium income group was more defaulted and Low income group is the least.
- For other application status more or less all the income groups are equally defaulted.

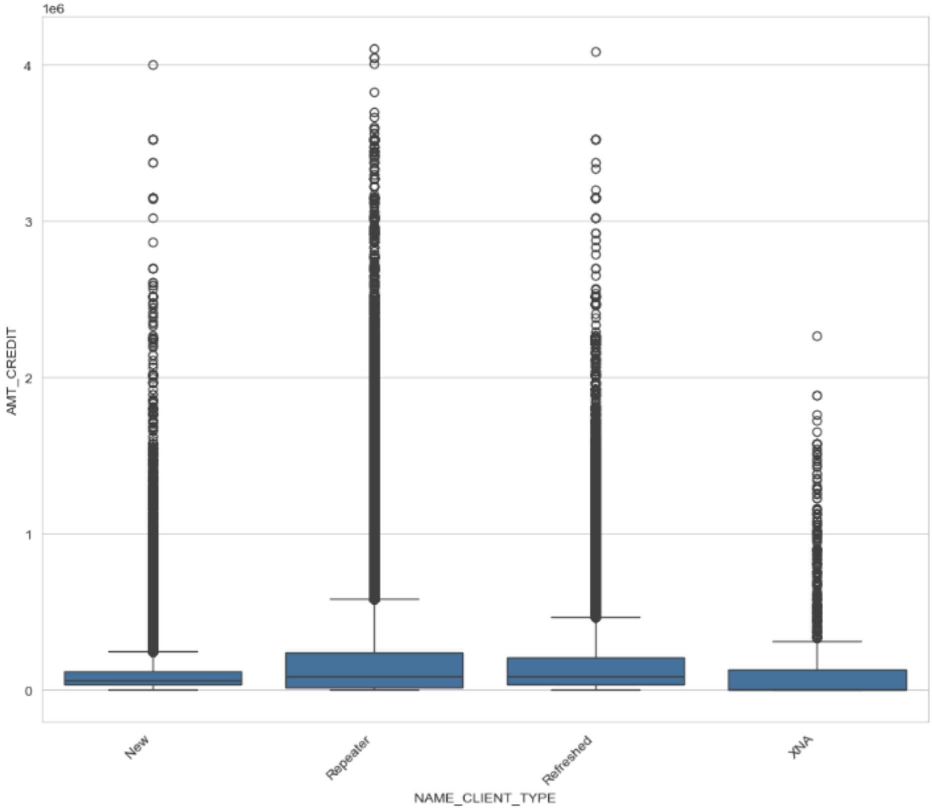
Analysis of Name_Contract_Status, Name_Portfolio and Target



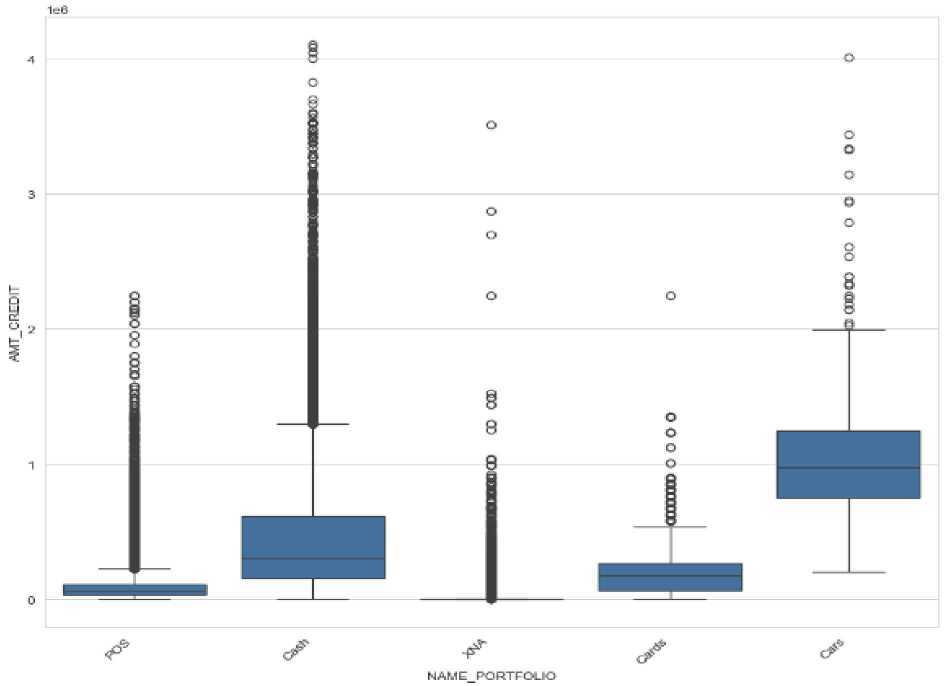
- Most of the clients were defaulted, who previously applied loan for Cards.
- For approved loan status the clients applied for Cars are less defaulted.
- For Refused loan status the clients applied for POS are less defaulted.

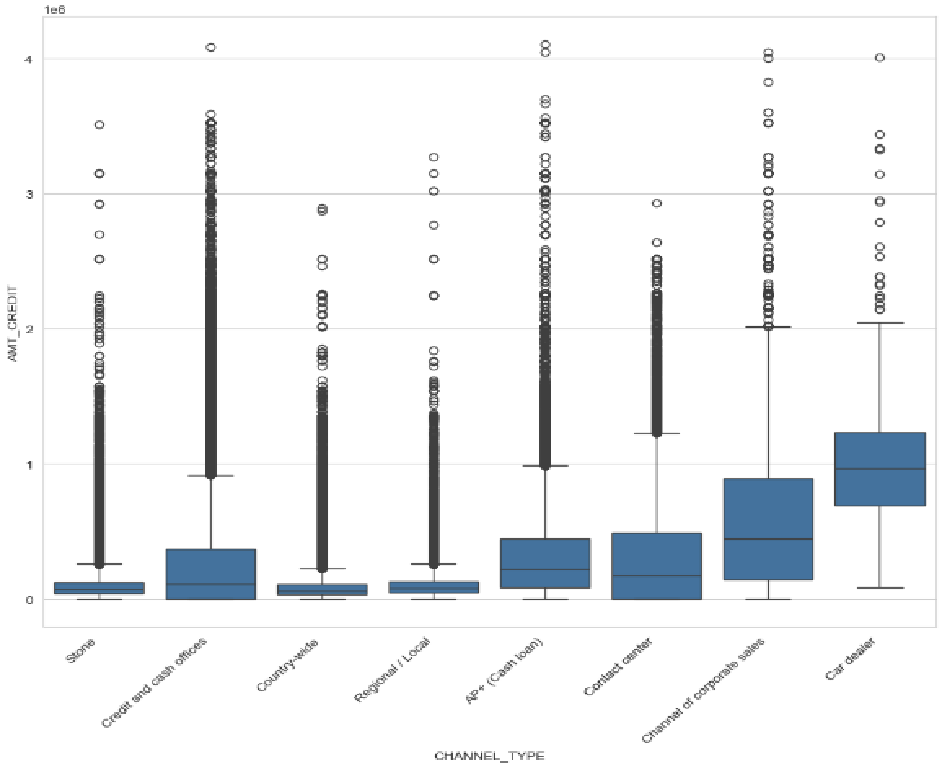
Analysis of Credit amount of Loan of various Category





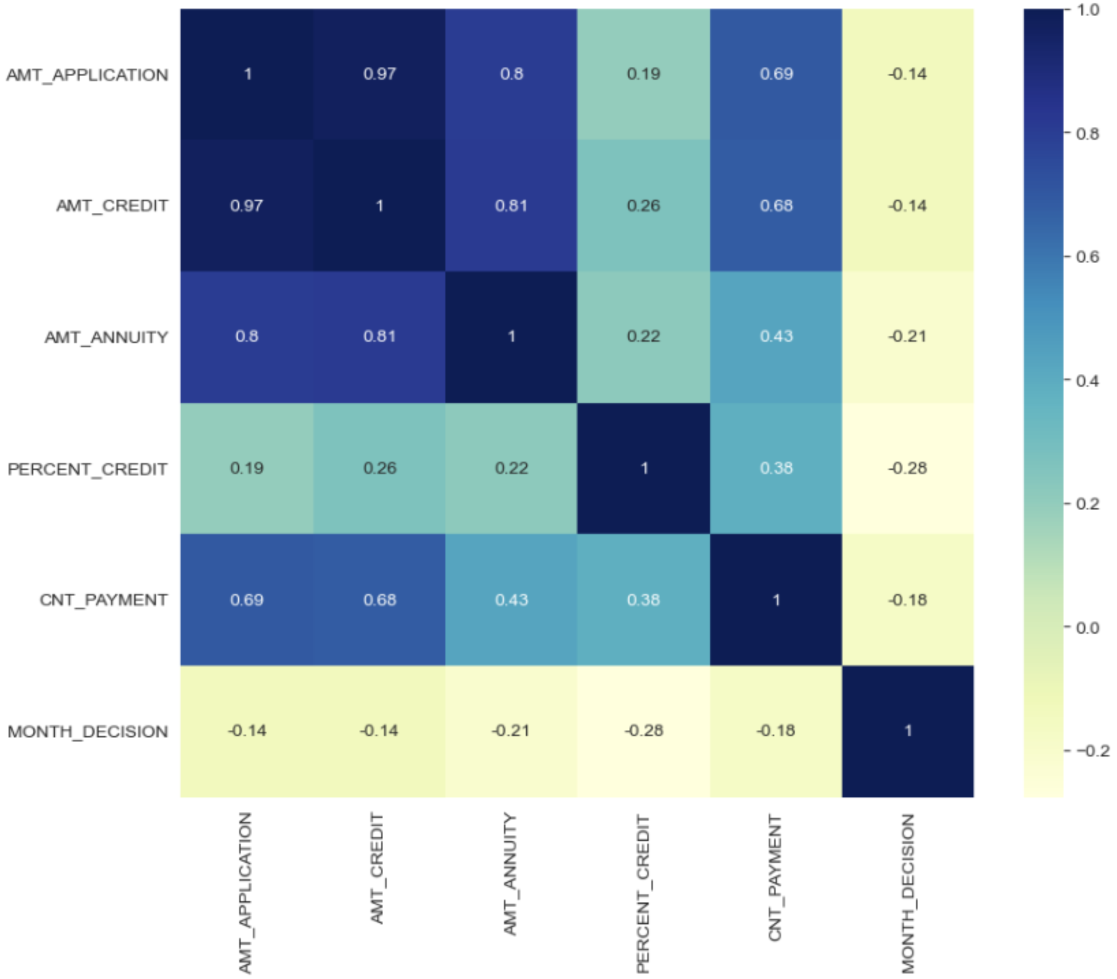
Analysis of Credit amount of Loan of various Category





- Cash loans are more credited in amount than Revolving and Consumer loans.
- Repeater clients get more amount loan than new and refreshed clients.
- The loan with portfolio Cars are more amount credited followed by Cash.
- The credit amount of the loan is more from the application channel type as car dealer followed by Channel of corporate sales,
- Credit and cash offices and Contact center. The amount is very less for Regional, Stone and Country-wide channels.

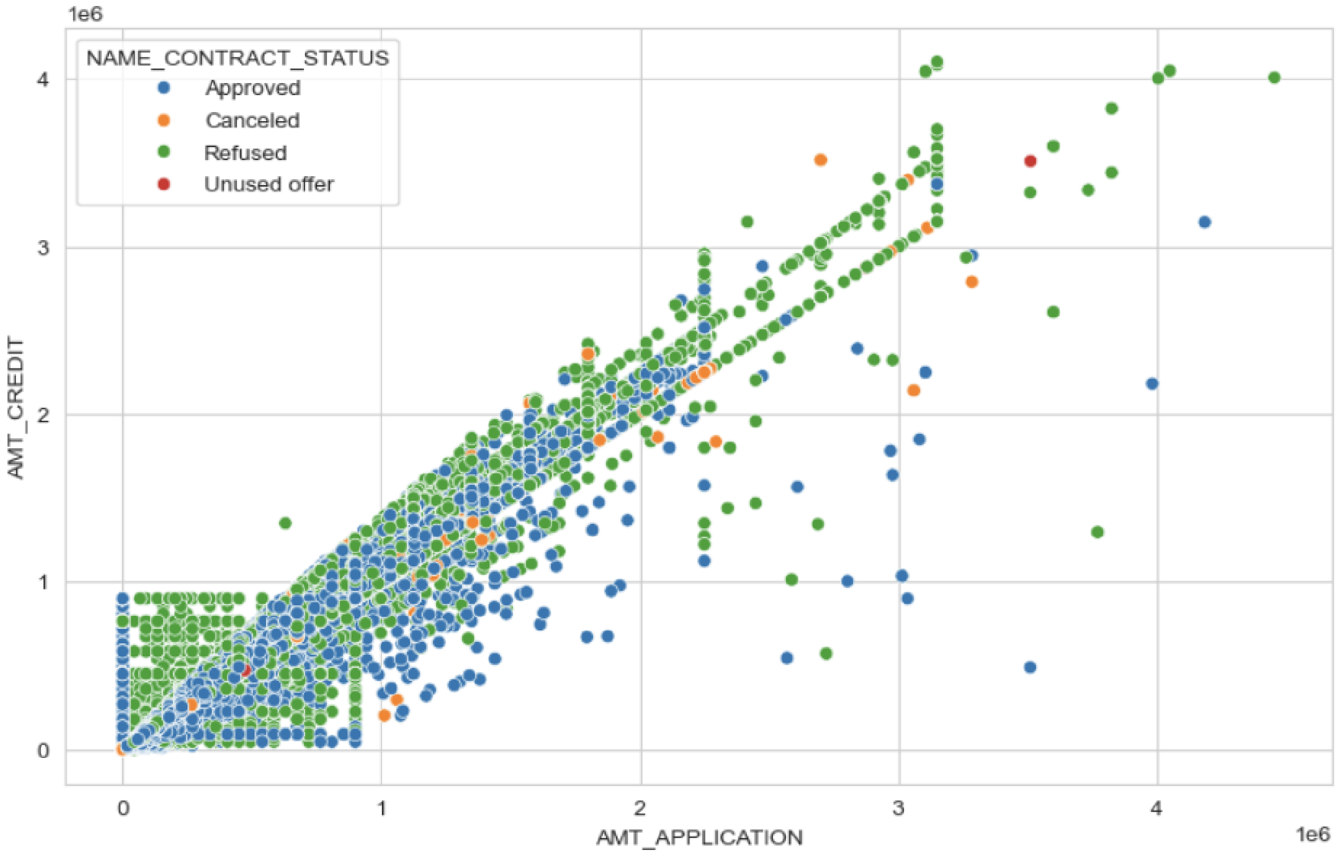
Corelation On Merged Data Set



	AMT_APPLICATION	AMT_CREDIT	AMT_ANNUITY	PERCENT_CREDIT	CNT_PAYMENT	MONTH_DECISION
AMT_APPLICATION	1	0.97	0.8	0.19	0.69	-0.14
AMT_CREDIT	0.97	1	0.81	0.26	0.68	-0.14
AMT_ANNUITY	0.8	0.81	1	0.22	0.43	-0.21
PERCENT_CREDIT	0.19	0.26	0.22	1	0.38	-0.28
CNT_PAYMENT	0.69	0.68	0.43	0.38	1	-0.18
MONTH_DECISION	-0.14	-0.14	-0.21	-0.28	-0.18	1

Highly corelate columns AMT_APPLICATION and AMT_CREDIT 0.97 AMT_APPLICATION and AMT_ANNUITY 0.8 AMT_CREDIT and AMT_ANNUITY 0.81 Moderately corelated columns AMT_APPLICATION and CNT_PAYMENT 0.69 AMT_CREDIT and CNT_PAYMENT 0.68

Analysis of AMT_Application, AMT_Credit, Name_Contract_Status



We can see that the applications are more concentrated on the lesser amount and so as the credited amount. Also, the credited amount is increased with respect to the application amount.

Conclusion

Highly recommended groups

- Clients whose previous loans were approved
- Highly educated clients
- Pensioners, State Servants followed by working and Commercial associates
- Senior citizens in all categories
- Clients with higher income
- Females are comparatively favourable than male.

Highly risky groups

- Clients with previously refused, cancelled or unused offers
- Low-income groups with previously refused status
- Unemployed and maternity leave groups
- Young clients are comparatively riskier than mid age clients and senior citizens
- Lower secondary and secondary educated clients
- Young low-income groups

Thank you